Title of the Invention

NUCLEOTIDE AND DEDUCED AMINO ACID SEQUENCES OF THE ENVELOPE 1 AND CORE GENES OF ISOLATES OF HEPATITIS C VIRUS AND THE USE OF REAGENTS DERIVED FROM THESE SEQUENCES IN DIAGNOSTIC METHODS AND VACCINES

The present application is a divisional application of pending U.S. Application Serial No. 08/290,665, filed August 15, 1994, which is a continuation-in-part of U.S. Application Serial No. 08/086,428, filed on June 29, 1993, now U.S. Patent No. 5,514,539.

Field Of Invention

The present invention is in the field of hepatitis virology. The invention relates to the complete nucleotide and deduced amino acid sequences of the envelope 1 (E1) and core genes of hepatitis C virus (HCV) isolates from around the world and the grouping of these isolates into fourteen distinct HCV genotypes. More specifically, this invention relates to oligonucleotides, peptides and recombinant proteins derived from the envelope 1 and core gene sequences of these isolates of hepatitis C virus and to diagnostic methods and vaccines which employ these reagents.

Background Of Invention

Hepatitis C, originally called non-A, non-B hepatitis, was first described in 1975 as a disease serologically distinct from hepatitis A and hepatitis B (Feinstone, S.M. et al. (1975) N. Engl. J. Med. 292:767-770). Although hepatitis C was (and is) the leading type of transfusion-associated hepatitis as well as an important part of community-acquired hepatitis, little progress was made in understanding the disease until the recent identification of hepatitis C virus (HCV) as the causative agent of hepatitis C via the cloning and sequencing of the

HCV genome (Choo, A.L. et al. (1989) Science 288:359-362). The sequence information generated by this study resulted in the characterization of HCV as a small, enveloped, positive-stranded RNA virus and led to the demonstration that HCV is a major cause of both acute and chronic hepatitis worldwide (Weiner, A.J. et al. (1990) Lancet 335:1-3). These observations, combined with studies showing that over 50% of acute cases of hepatitis C progress to chronicity with 20% of these resulting in cirrhosis and an undetermined proportion progressing to liver cancer, have led to tremendous efforts by investigators within the hepatitis C field to develop diagnostic assays and vaccines which can detect and prevent hepatitis C infection.

The cloning and sequencing of the HCV genome by Choo et al. (1989) has permitted the development of serologic tests which can detect HCV or antibody to HCV (Kuo, G. et al. (1989) Science 244:362-364). In addition, the work of Choo et al. has also allowed the development of methods for detecting HCV infection via amplification of HCV RNA sequences by reverse transcription and cDNA polymerase chain reaction (RT-PCR) using primers derived from the HCV genomic sequence (Weiner, A.J. et al.). However, although the development of these diagnostic methods has resulted in improved diagnosis of HCV infection, only approximately 60% of cases of hepatitis C are associated with a factor identified as contributing to transmission of HCV (Alter, M.J. et al. (1989) JAMA 262:1201-1205). This observation suggests that effective control of hepatitis C transmission is likely to occur only via universal pediatric vaccination as has been initiated recently for hepatitis B virus. Unfortunately, attempts to date to protect chimpanzees from hepatitis C infection via administration of recombinant vaccines have had only limited success. Moreover, the apparent genetic heterogeneity of HCV, as indicated by the recent assignment

of all available HCV isolates to one of four genotypes, I-IV (Okamoto, H. et al. (1992) J. Gen. Virol; 73:673-679), presents additional hurdles which must be overcome in order to develop accurate and effective diagnostic assays and vaccines.

For example, one possible obstacle to the development of effective hepatitis C vaccines would arise if the observed genetic heterogeneity of HCV reflects serologic heterogeneity. In such a case, the most genetically diverse strains of HCV may then represent different serotypes of HCV with the result being that infection with one strain may not protect against infection with another. Indeed, the inability of one strain to protect against infection with another strain was recently noted by both Farci et al. (Farci, P. et al. (1992) Science 258:135-140) and Prince et al. (Prince, A.M. et al. (1992) J. Infect. Dis. 165:438-443), each of whom presented evidence that while infection with one strain of HCV does modify the degree of the hepatitis C associated with the reinfection, it does not protect against reinfection with a closely related strain. The genetic heterogeneity among different HCV strains also increases the difficulty encountered in developing RT-PCR assays to detect HCV infection since such heterogeneity often results in false-negative results because of primer and template mismatch. In addition, currently used serologic tests for detection of HCV or for detection of antibody to HCV are not sufficiently well developed to detect all of the HCV genotypes which might exist in a given blood sample. Finally, in terms of choosing the proper treatment modality to combat hepatitis infection, the inability of presently available serologic assays to distinguish among the various genotypes of HCV represents a significant shortcoming in that recent reports suggest that an HCV-infected patient's response to therapy might be related to the genotype of the infectious virus (Yoshioka, K. et al. (1992) Hepatology

16:293-299; Kanai, K. et al. (1992) Lancet 339:1543; Lan, J.Y.N. et al. (1992) Hepatology 16:209A). Indeed, the data presented in the above studies suggest that the closely related genotypes I and II are less responsive to interferon therapy than are the closely related genotypes III and IV. Moreover, preliminary data by Pozzato et al. (Pozzato, G. et al. (1991) Lancet 338:509) suggests that different genotypes may be associated with different types or degrees of clinical disease. Taken together, these studies suggest that before effective vaccines against HCV infection can be developed, and indeed, before more accurate and effective methods for diagnosis and treatment of HCV infection can be produced, one must obtain a greater knowledge about the genetic and serologic diversity of HCV isolates.

In a recent attempt to gain an understanding of the extent of genetic heterogeneity among HCV strains, Bukh et al. carried out a detailed analysis of HCV isolates via the use of PCR technology to amplify different regions of the HCV genome (Bukh, J. et al. (1992a) Proc. Natl. Acad. Sci. 89:187-191). Following PCR amplification, the 5'-noncoding (5' NC) portion of the genomes of various HCV isolates were sequenced and it was found that primer pairs designed from conserved regions of the 5' NC region of the HCV genome were more sensitive for detecting the presence of HCV than were primer pairs representing other portions of the genome (Bukh, J. et al. (1992b) Proc. Natl. Acad. Sci. U.S.A. 89:4942-4946). In addition, the authors noted that although many of the HCV isolates examined could be classified into the four genotypes described by Okamoto et al. (1992), other previously undescribed genotypes emerged based on genetic heterogeneity observed in the 5' NC region of the various isolates. One of the most prominent of these newly noted genotypes comprised a group of related viruses that contained the most genetically divergent 5' NC regions of those studied. This group of viruses,

tentatively classified as a fifth genotype, are very similar to strains recently described by others (Cha, T.-A et al. (1992) Proc. Natl. Acad. Sci. U.S.A. 89:7144-7148; Chan, S-W. et al. (1992) J. Gen. Virol., 73:1131-1141 and Lee, C-H et al. (1992) J. Clin. Microbio. 30:1602-1604). In addition, at least four more putative genotypes were identified thereby providing evidence that the genetic heterogeneity of HCV was more extensive than previously appreciated.

However, while the studies of Bukh et al. (1992a and b) provided new and useful information on the genetic heterogeneity of HCV, it is widely appreciated by those skilled in the art that the three structural genes of HCV, core (C), envelope (E1) and envelope 2/nonstructural 1 (E2/NS1) are the most important for the development of serologic diagnostics and vaccines since it is the product of these genes that constitutes the hepatitis C virion. Thus, a determination of the nucleotide sequence of one or all of the structural genes of a variety of HCV isolates would be useful in designing reagents for use in diagnostic assays and vaccines since a demonstration of genetic heterogeneity in a structural gene(s) of HCV isolates might suggest that some of the HCV genotypes represent distinct serotypes of HCV based upon the previously observed relationship between genetic heterogeneity and serologic heterogeneity among another group of single-stranded, positive-sense RNA viruses, the picornaviruses (Ruechert, R.R. "Picornaviridae and their replication", in Fields, B.N. et al., eds. Virology, New York: Raven Press, Ltd. (1990) 507-548).

Summary of Invention

The present invention relates to cDNAs encoding the complete nucleotide sequence of either the envelope 1 (E1) gene or the core (C) gene of an isolate of human hepatitis C virus (HCV).

The present invention also relates to the nucleic acid and deduced amino acid sequences of these E1 and core cDNAs.

It is an object of this invention to provide synthetic nucleic acid sequences capable of directing production of recombinant E1 and core proteins, as well as equivalent natural nucleic acid sequences. Such natural nucleic acid sequences may be isolated from a cDNA or genomic library from which the gene capable of directing synthesis of the E1 or core proteins may be identified and isolated. For purposes of this application, nucleic acid sequence refers to RNA, DNA, cDNA or any synthetic variant thereof which encodes for peptides.

The invention also relates to the method of preparing recombinant E1 and core proteins derived from E1 and core cDNA sequences respectively by cloning the nucleic acid encoding either the recombinant E1 or core protein and inserting the cDNA into an expression vector and expressing the recombinant protein in a host cell.

The invention also relates to isolated and substantially purified recombinant E1 and core proteins and analogs thereof encoded by E1 and core cDNAs respectively.

The invention further relates to the use of recombinant E1 and core proteins, either alone, or in combination with each other, as diagnostic agents and as vaccines.

The present invention also relates to the recombinant production of the core protein of the present invention to contain a second protein on its surface and therefore serve as a carrier in a multivalent vaccine preparation. Further, the present invention relates to the use of the self aggregating core or envelope proteins as a drug delivery system for anti-virals.

The invention also relates to the use of single-stranded antisense poly- or oligonucleotides derived from E1 or core cDNAs, or from both E1 and core cDNAs, to

inhibit expression of hepatitis C E1 and/or core genes.

The invention further relates to multiple computer-generated alignments of the nucleotide and deduced amino acid sequences of the E1 and core cDNAs. These multiple sequence alignments produce consensus sequences which serve to highlight regions of homology and non-homology between sequences found within the same genotype or in different genotypes and hence, these alignments can be used by one skilled in the art to design peptides and oligonucleotides useful as reagents in diagnostic assays and vaccines.

The invention therefore also relates to purified and isolated peptides and analogs thereof derived from E1 and core cDNA sequences.

The invention further relates to the use of these peptides as diagnostic agents and vaccines.

The present invention also encompasses methods of detecting antibodies specific for hepatitis C virus in biological samples. The methods of detecting HCV or antibodies to HCV disclosed in the present invention are useful for diagnosis of infection and disease caused by HCV and for monitoring the progression of such disease. Such methods are also useful for monitoring the efficacy of therapeutic agents during the course of treatment of HCV infection and disease in a mammal.

The invention also provides a kit for the detection of antibodies specific for HCV in a biological sample where said kit contains at least one purified and isolated peptide derived from the E1 or core cDNA sequences. In addition, the invention provides for a kit containing at least one purified and isolated peptide derived from the E1 cDNA sequences and at least one purified and isolated peptide derived from the core cDNA sequences.

The invention further provides isolated and purified genotype-specific oligonucleotides and analogs

thereof derived from E1 and core cDNA sequences.

The invention also relates to methods for detecting the presence of hepatitis C virus in a mammal, said methods comprising analyzing the RNA of a mammal for the presence of hepatitis C virus. The invention further relates to methods for determining the genotype of hepatitis C virus present in a mammal. This method is useful in determining the proper course of treatment for an HCV-infected patient.

The invention also provides a diagnostic kit for the detection of hepatitis C virus in a biological sample. The kit comprises purified and isolated nucleic acid sequences useful as primers for reverse-transcription polymerase chain reaction (RT-PCR) analysis of RNA for the presence of hepatitis C virus genomic RNA.

The invention further provides a diagnostic kit for the determination of the genotype of a hepatitis C virus present in a mammal. The kit comprises purified and isolated nucleic acid sequences useful as primers for RT-PCR analysis of RNA for the presence of HCV in a biological sample and purified and isolated nucleic acid sequences useful as hybridization probes in determining the genotype of the HCV isolate detected in PCR analysis.

This invention also relates to pharmaceutical compositions useful in prevention or treatment of hepatitis C in a mammal.

Description of Figures

Figures 1 A-H show computer generated sequence alignments of the nucleotide sequences of 51 HCV E1 cDNAs. The single letter abbreviations used for the nucleotides shown in Figures 1A-H are those standardly used in the art. Figure 1A shows the alignment of SEQ ID NOs:1-8 to produce a consensus sequence for genotype I/1a. Figure 1B shows the alignment of SEQ ID NOs:9-25 to produce a consensus sequence for genotype II/1b. Figure 1C shows the alignment

of SEQ ID NOs:26-29 to produce a consensus sequence for genotype III/2a. Figure 1D shows the alignment of SEQ ID NOs:30-33 to produce a consensus sequence for genotype IV/2b. Figure 1E shows the alignment of SEQ ID NOs:35-39 to produce a consensus sequence for genotype V/3a. Figure 1F shows the computer alignment of SEQ ID NOs:42-43 to produce a "consensus" sequence for genotype 4C where the "consensus" sequence given is that of SEQ ID NO:42. Figure 1G shows the alignment of SEQ ID NOs:45-50 to produce a consensus sequence for genotype 5a. The nucleotides shown in capital letters in the consensus sequences of Figures 1A-G are those conserved within a genotype while nucleotides shown in lower case letters in the consensus sequences are those variable within a genotype. In addition, in Figures 1A-E and 1G, when the lower case letter is shown in a consensus sequence, the lower case letter represents the nucleotide found most frequently in the sequences aligned to produce the consensus sequence. In Figure 1F, the lower case letters shown in the consensus sequence are nucleotides in SEQ ID NO:42 which differ from nucleotides found in the same positions in SEQ ID NO:43. Finally, a hyphen at a nucleotide position in the consensus sequences in Figures 1A-G indicates that two nucleotides were found in equal numbers at that position in the aligned sequences. In the aligned sequences, nucleotides are shown in lower case letters if they differed from the nucleotides of both adjacent isolates. Figure 1H shows the alignment of the consensus sequences of Figures 1A-G with SEQ ID NO:34 (genotype 2c), SEQ ID NO:40 (genotype 4a), SEQ ID NO:41 (genotype 4b), SEQ ID NO:44 (genotype 4d) and SEQ ID NO:51 (genotype 6a) to produce a consensus sequence for all twelve genotypes. This consensus sequence is shown as the bottom line of Figure 1H where the nucleotides shown in capital letters are conserved among all genotypes and a blank space indicates that the nucleotide at that position is not conserved among all genotypes.

Figures 2A-H show computer alignments of the deduced amino acid sequences of 51 HCV E1 cDNAs. The single letter abbreviations used for the amino acids shown in Figures 2A-H follow the conventional amino acid shorthand for the twenty naturally occurring amino acids. Figure 2A shows the alignment of SEQ ID NOs:52-59 to produce a consensus sequence for genotype I/1a. Figure 2B shows the alignment of SEQ ID NOs:60-76 to produce a consensus sequence for genotype II/1b. Figure 2C shows the alignment of SEQ ID NOs:77-80 to produce a consensus sequence for genotype III/2a. Figure 2D shows the alignment of SEQ ID NOs:81-84 to produce a consensus sequence for genotype IV/2b. Figure 2E shows the alignment of SEQ ID NOs:86-90 to produce a consensus sequence for genotype V/3a. Figure 2F shows the computer alignment of SEQ ID NOs:93-94 to produce a consensus sequence for genotype 4c. Figure 2G shows the alignment of SEQ ID NOs:96-101 to produce a consensus sequence for genotype 5a. The amino acids shown in capital letters in the consensus sequences of Figures 2A-G are those conserved within a genotype while amino acids shown in lower case letters in the consensus sequences are those variable within a genotype. In addition, in Figures 2A-E and 2G when the lower case letter is shown in a consensus sequence, the letter represents the amino acid found most frequently in the sequences aligned to produce the consensus sequence. In Figure 2F, the lower case letters shown in the consensus sequence are amino acids in SEQ ID NO:93 which differ from amino acids found in the same positions in SEQ ID NO:94. Finally, a hyphen at an amino acid position in the consensus sequences of Figures 2A-G indicates that two amino acids were found in equal numbers at that position in the aligned sequences. In the aligned sequences, amino acids are shown in lower case letters if they differed from the amino acids of both adjacent isolates. Figure 2H shows the alignment of the consensus sequences of Figures 2A-G

with SEQ ID NO:85 (genotype 2c), SEQ ID NO:91 (genotype 4a), SEQ ID NO:92 (genotype 4b), SEQ ID NO:95 (genotype 4d) and SEQ ID NO:102 (genotype 6a) to produce a consensus sequence for all twelve genotypes. This consensus sequence is shown as the bottom line of Figure 2H where the amino acids shown in capital letters are conserved among all genotypes and a blank space indicates that the amino acid at that position is not conserved among all genotypes.

Figure 3 shows multiple sequence alignment of the deduced amino acid sequence of the E1 gene of 51 HCV isolates collected worldwide. The consensus sequence of the E1 protein is shown in boldface (top). In the consensus sequence cysteine residues are highlighted with stars, potential N-linked glycosylation sites are underlined, and invariant amino acids are capitalized, whereas variable amino acids are shown in lower case letters. In the alignment, amino acids are shown in lower case letters if they differed from the amino acid of both adjacent isolates. Amino acid residues shown in bold print in the alignment represent residues which at that position in the amino acid sequence are genotype-specific. Amino acids that were invariant among all HCV isolates are shown as hyphens (-) in the alignment. Amino acid positions correspond to those of the HCV prototype sequence (HCV-1, Choo, L. et al. (1991) Proc. Natl. Acad. Sci. USA 88:2451-2455) with the first amino acid of the E1 protein at position 192. The grouping of isolates into 12 genotypes (I/1a, II/1b, III/2a, IV/2b, V/3a, 2c, 4a, 4b, 4c, 4d, 5a and 6a) is indicated.

Figure 4 shows a dendrogram of the genetic relatedness of the twelve genotypes of HCV based on the percent amino acid identity of the E1 gene of the HCV genome. The twelve genotypes shown are designated as I/1a, II/1b, III/2a, IV/2b, V/3a, 2c, 4a, 4b, 4c, 4d, 5a and 6a. The shaded bars represent a range showing the maximum and minimum homology between the amino acid sequence of any one

isolate of the genotype indicated and the amino acid sequence of any other isolate.

Figure 5 shows the distribution of the complete E1 gene sequence of 74 HCV isolates into the twelve HCV genotypes in the 12 countries studied. For 51 of these HCV isolates, including 8 isolates of genotype I/1a, 17 isolates of genotype II/1b and 26 isolates comprising the additional 10 genotypes, the complete E1 gene sequence was determined. In the remaining 23 isolates, all of genotypes I/1a and II/1b, the genotype assignment was based on only a partial E1 gene sequence. The partially sequenced isolates did not represent additional genotypes in any of the 12 countries. The number of isolates of a particular genotype is given in each of the 12 countries studied. For ease of viewing, those genotypes designated by two terms (e.g., I/1a) are indicated by the latter term (e.g. 1a). The designations used for each country are: Denmark (DK); Dominican Republic (DR); Germany (D); Hong Kong (HK); India (IND); Sardinia, Italy (S); Peru (P); South Africa (SA); Sweden (SW); Taiwan (T); United States (US); and Zaire (Z). National borders depicted in this figure represent those existing at the time of sampling.

Figures 6A-K show computer generated sequence alignments of the nucleotide sequences of 52 HCV core cDNAs. Single letter abbreviations used for the nucleotides shown in Figures 6A-J are those standardly used in the art. Figure 6A shows the alignment of SEQ ID NOs: 103-108 to produce a consensus sequence for genotype I/Ia. Figure 6B shows the alignment of SEQ ID NOs: 109-124 to produce a consensus sequence for genotype II/1b. Figure 6C shows the alignments of the sequences comprising minor genotypes I/1a (SEQ ID NOS: 103-108) and II/1b (SEQ ID NOs: 109-124) to produce a consensus sequence for the major genotype, genotype 1. Figure 6D shows the alignment of SEQ ID NOs: 125-128 to produce a consensus sequence for genotype III/2a. Figure 6E shows the alignment of SEQ ID

NOs: 129-133 to produce a consensus sequence for genotype IV/2b. Figure 6F shows the alignment of the sequences of minor genotypes III/2a (SEQ ID NOs: 125-128), IV/2b (SEQ ID NOs: 129-133) and 2c (SEQ ID NO: 134) to produce a consensus sequence for the major genotype, genotype 2. Figure 6G shows the alignment of SEQ ID NOs: 135-138 to produce a consensus sequence for genotype V/3a. Figure 6H shows the computer alignment of the sequences of minor genotypes 4a-4f (SEQ ID NOs: 139-145) to produce a consensus sequence for the major genotype, genotype 4. Figure 6I shows the alignment of SEQ ID NOs: 146-153 to produce a consensus sequence for genotype 5a. The nucleotides shown in capital letters in the consensus sequences in Figure 6A-I are those conserved within the genotype while nucleotides shown in lower case letters in the consensus sequences are those variable within a genotype. In addition, when the lower case letter is shown in the consensus sequence, the lower case letter represents the nucleotide found most frequently in the sequences aligned to produce that consensus sequence. Moreover, a hyphen at a nucleotide position in the consensus sequences in Figures 6A-6I indicates that two nucleotides were found in equal numbers at that position in the sequences aligned to produce the consensus sequence. Finally, nucleotides are shown in lower case letters in the sequences aligned to produce each consensus sequence shown in Figures 6A-6I, if they differed from the nucleotides of both adjacent isolates. Figure 6J shows the alignment of the consensus sequences of major genotypes 1 (Figure 6C), 2 (Figure 6F), 3 (Figure 6G), 4 (Figure 6H), 5 (Figure 6I) and 6 (SEQ ID NO: 154) to produce a consensus sequence for all genotypes and Figure 6K shows the alignment of consensus sequences of Figures 6A, 6B, 6D, 6E, 6G and 6I with SEQ ID NO:134 (genotype 2c), SEQ ID NO:139 (genotype 4a), SEQ ID NO:141 (genotype 4b), SEQ ID NO:143 (genotype 4c), SEQ ID NO:145 (genotype 4d), SEQ ID NO:142 (genotype 4e), SEQ ID NO:140

(genotype 4f) and SEQ ID NO:154 (genotype 6a) to produce a consensus sequence for all fourteen genotypes. The nucleotides shown in capital letters in the consensus sequences of Figures 6J and 6K are conserved among all genotypes and the nucleotide shown in lower case letter represent the nucleotides found most frequently in the sequences aligned to produce this consensus sequence. In addition, the presence of a hyphen at a nucleotide position in all fourteen sequences aligned in Figure 6K indicates that the nucleotide found at that position in the aligned sequences is the same as nucleotide shown at the corresponding position in the consensus sequences of Figure 6K.

Figures 7A-7J show computer alignments of the deduced amino acid sequences of the 52 HCV core cDNAs. The single letter abbreviations used for the amino acids shown in Figures 7A-7J follow the conventional amino acid short hand for the twenty natural occurring amino acids. Figure 7A shows the alignment of SEQ ID NOs: 155-160 to produce a consensus sequence for genotype I/1a. Figure 7B shows the alignment of SEQ ID NOs: 161-176 to produce a consensus sequence for genotype II/1b. Figure 7C shows the alignment of the sequences comprising minor genotypes I/a (SEQ ID NOS: 155-160) and II/1b (SEQ ID NOS: 161-176) to produce a consensus sequence for the major genotype, genotype 1. Figure 7D shows the alignment of SEQ ID NOs: 177-180 to produce a consensus sequence for genotype III/2a. Figure 7E shows the alignment of SEQ ID NOs: 181-185 to produce a consensus sequence for genotype IV/2b. Figure 7F shows the alignment of the sequences of minor genotypes III/2a (SEQ ID NOS: 177-180), IV/2b (SEQ ID NOS: 181-185) and 2c (SEQ ID NO: 186) to produce a consensus sequence for the major genotype, genotype 2. Figure 7G shows the alignment of SEQ ID NOs: 187-190 to produce a consensus sequence for genotype V/3a. Figure 7H shows the computer alignment of the sequences of minor genotypes 4a-4f (SEQ ID NOs: 191-

197) to produce a consensus sequence for the major genotype, genotype 4. Figure 7I shows the alignment of SEQ ID NOs: 198-205 to produce a consensus sequence for genotype 5a. The amino acids shown in capital letters in the consensus sequences of Figures 7A-7I are those conserved within the genotype while amino acids shown in lower case letters in the consensus sequences are those variable within the genotype. In addition, when a lower case letter is found in the consensus sequences shown in Figures 7A-7I, the letter represents the amino acid found most frequently in the sequences aligned to produce that consensus sequence. Moreover, a hyphen in an amino acid position in the consensus sequences of Figures 7A-7I indicates that two amino acids were found in equal numbers at that position in the sequences aligned to produce that consensus sequence. Finally, amino acids are shown in lower case letters in the sequences aligned to produce the consensus sequences shown in Figures 7A-7I if these amino acids differed from the amino acids of both adjacent isolates. Figure 7J shows the alignment of the consensus sequences of major genotypes 1 (Figure 7C), 2 (Figure 7F), 3 (Figure 7G), 4 (Figure 7H), 5 (Figure 7I) and 6 (SEQ ID NO: 154) to produce a consensus sequence for all genotypes and Figure 7K shows the alignment of the consensus sequences of Figures 7A, 7B, 7D, 7E, 7G and 7I with SEQ ID NO:186 (genotype 2c), SEQ ID NO:191 (genotype 4a), SEQ ID NO:193 (genotype 4b), SEQ ID NO:195 (genotype 4c), SEQ ID NO:197 (genotype 4d), SEQ ID NO:194 (genotype 4e), SEQ ID NO:192 (genotype 4f) and SEQ ID NO:206 (genotype 6a) to produce a consensus sequence for all fourteen genotypes. The amino acids shown in capital letters in the consensus sequences shown in Figures 7J and 7K are conserved among all genotypes while the amino acids shown in lower case letters represent amino acids found most frequently in the sequences aligned to produce this consensus sequence. In addition, the presence of a hyphen at an amino acid

position in all fourteen sequences aligned in Figure 7K indicates that the amino acid found at that position in the aligned sequences is the same as the amino acid shown at the corresponding position in the consensus sequence of Figure 7K.

Figure 8 shows phylogenetic trees illustrating the calculated evolutionary relationships of the different HCV isolates based upon the C gene sequence of 52 HCV isolates and the E1 gene sequence of 51 HCV isolates, respectively. The phylogenetic trees were constructed by the unweighted pair-group method with arithmetic mean (Nei, M. (1987) *Molecular Evolutionary Genetics* (Columbia University Press, New York, N.Y.), pp 287-326) using the computer software package "Gene Works" from IntelliGenetics. The lengths of the horizontal lines connecting the sequences, given in absolute values from 0 to 1, are proportional to the estimated genetic distances between the sequences. Genotype designations of HCV isolates are indicated. In 45 HCV isolates, both the C and the E1 gene sequences were determined.

Detailed Description Of Invention

The present invention relates to cDNAs encoding the complete nucleotide sequence of the envelope 1 (E1) and core genes of isolates of human hepatitis C virus (HCV). The E1 cDNAs of the present invention were obtained as follows. Viral RNA was extracted from serum collected from humans infected with hepatitis C virus and the viral RNA was then reverse transcribed and amplified by polymerase chain reaction using primers deduced from the sequence of the HCV strain H-77 (Ogata, N. et al. (1991) Proc. Natl. Acad. Sci. U.S.A. 88:3392-3396). The amplified cDNA was then isolated by gel electrophoresis and sequenced.

The present invention further relates to the nucleotide sequences of the cDNAs encoding the E1 gene of 51 HCV isolates. These nucleotide sequences are shown in

the sequence listing as SEQ ID NO:1 through SEQ ID NO:51.

The abbreviations used for the nucleotides are those standardly used in the art.

The deduced amino acid sequence of each of SEQ ID NO:1 through SEQ ID NO:51 are presented in the sequence listing as SEQ ID NO:52 through SEQ ID NO:102 where the amino acid sequence in SEQ ID NO:52 is deduced from the nucleotide sequence shown in SEQ ID NO:1, the amino acid sequence shown in SEQ ID NO:53 is deduced from the nucleotide sequence shown in SEQ ID NO:2 and so on. The deduced amino acid sequence of each of SEQ ID Nos:52-102 starts at nucleotide 1 of the corresponding nucleic acid sequence shown in SEQ ID NOs:1-51 and extends 575 nucleotides to a total length of 576 nucleotides.

The three letter abbreviations used in SEQ ID Nos:52-102 follow the conventional amino acid shorthand for the twenty naturally occurring amino acids.

The present invention also relates to the nucleotide sequences of the cDNAs encoding the core gene of 52 HCV isolates. These nucleotide sequences are shown in the sequence listing as SEQ ID NO:103 through SEQ ID NO:154.

The core cDNAs of the present invention were obtained as follows. Viral RNA was extracted from serum and reversed transcribed as described above for cloning of the E1 cDNAs. The core cDNAs of the present invention were then amplified by polymerase chain reaction using primers deduced from previously determined sequences that flank the core gene (Bukh et al. (1992)) Proc. Natl. Acad. Sci. U.S.A., 89: 4942-4946; Bukh et al. (1993) Proc. Natl. Acad. Sci. U.S.A., 90: 8234-8238).

The deduced amino acid sequence of each of SEQ ID NO:103 through SEQ ID NO:154 are presented in the sequence listing as SEQ ID NO:155 through SEQ ID NO:206 where the amino acid sequence in SEQ ID NO:155 is deduced from the nucleotide sequence shown in SEQ ID NO:103, the amino acid

sequence shown in SEQ ID NO:156 is deduced from the nucleotide sequence shown in SEQ ID NO:104 and so on. The deduced amino acid sequence of each of SEQ ID NOs: 155-206 starts at nucleotide 1 of the corresponding nucleotide sequence shown in SEQ ID NOs:103-154 and extends 572 nucleotides to a total length of 573 nucleotides.

Preferably, the E1 and core proteins and peptides of the present invention are substantially homologous to, and most preferably biologically equivalent to, native HCV E1 and core proteins and peptides. By "biologically equivalent" as used throughout the specification and claims, it is meant that the compositions are immunogenically equivalent to the native E1 and core proteins and peptides. The E1 and core proteins and peptides of the present invention may also stimulate the production of protective antibodies upon injection into a mammal that would serve to protect the mammal upon challenge with HCV. By "substantially homologous" as used throughout the ensuing specification and claims to describe E1 and core proteins and peptides, it is meant a degree of homology in the amino acid sequence of the E1 and core proteins and peptides to the native E1 and core proteins and peptides respectively. Preferably the degree of homology is in excess of 90, preferably in excess of 95, with a particularly preferred group of proteins being in excess of 99 homologous with the native E1 or core proteins and peptides.

Variations are contemplated in the cDNA sequences shown in SEQ ID NO:1 through SEQ ID NO:51 and in SEQ ID NO:103 through SEQ ID NO:154 which will result in a nucleic acid sequence that is capable of directing production of analogs of the corresponding protein shown in SEQ ID NO:52 through SEQ ID NO:102 and in SEQ ID NO:155 through SEQ ID NO:206. It should be noted that the cDNA sequences set forth above represent a preferred embodiment of the present invention. Due to the degeneracy of the genetic code, it

is to be understood that numerous choices of nucleotides may be made that will lead to a DNA sequence capable of directing production of the instant protein or its analogs. As such, DNA sequences which are functionally equivalent to the sequence set forth above or which are functionally equivalent to sequences that would direct production of analogs of the E1 and core proteins produced pursuant to the amino acid sequences set forth above, are intended to be encompassed within the present invention.

The term analog as used throughout the specification or claims to describe the E1 and core proteins and peptides of the present invention, includes any protein or peptide having an amino acid residue sequence substantially identical to a sequence specifically shown herein in which one or more residues have been conservatively substituted with a biologically equivalent residue. Examples of conservative substitutions include the substitution of one polar (hydrophobic) residue such as isoleucine, valine, leucine or methionine for another, the substitution of one polar (hydrophilic) residue for another such as between arginine and lysine, between glutamine and asparagine, between glycine and serine, the substitution of one basic residue such as lysine, arginine or histidine for another, or the substitution of one acidic residue, such as aspartic acid or glutamic acid for another.

The phrase "conservative substitution" also includes the use of a chemically derivatized residue in place of a non-derivatized residue provided that the resulting protein or peptide is biologically equivalent to the native E1 or core protein or peptide.

"Chemical derivative" refers to an E1 or core protein or peptide having one or more residues chemically derivatized by reaction of a functional side group. Examples of such derivatized molecules, include but are not limited to, those molecules in which free amino groups have been derivatized to form amine hydrochlorides, p-toluene

sulfonyl groups, carbobenzoxy groups, t-butyloxycarbonyl groups, chloracetyl groups or formyl groups. Free carboxyl groups may be derivatized to form salts, methyl and ethyl esters or other types of esters or hydrazides. Free hydroxyl groups may be derivatized to form O-acyl or O-alkyl derivatives. The imidazole nitrogen of histidine may be derivatized to form N-imbenzylhistidine. Also included as chemical derivatives are those proteins or peptides which contain one or more naturally-occurring amino acid derivatives of the twenty standard amino acids. For examples: 4-hydroxyproline may be substituted for proline; 5-hydroxylysine may be substituted for lysine; 3-methylhistidine may be substituted for histidine; homoserine may be substituted for serine; and ornithine may be substituted for lysine. The E1 and core proteins and peptide of the present invention also includes any protein or peptide having one or more additions and/or deletions of residues relative to the sequence of a peptide whose sequence is shown herein, so long as the peptide is biologically equivalent to the native E1 or core protein or peptide.

The present invention also includes a recombinant DNA method for the manufacture of HCV E1 and core proteins. In this method, natural or synthetic nucleic acid sequences may be used to direct the production of E1 and core proteins.

In one embodiment of the invention, the method comprises:

(a) preparation of a nucleic acid sequence capable of directing a host organism to produce HCV E1 or core protein;

(b) cloning the nucleic acid sequence into a vector capable of being transferred into and replicated in a host organism, such vector containing operational elements for the nucleic acid sequence;

(c) transferring the vector containing the

nucleic acid and operational elements into a host organism capable of expressing the protein;

(d) culturing the host organism under conditions appropriate for amplification of the vector and expression of the protein; and

(e) harvesting the protein.

In another embodiment of the invention, the method for the recombinant DNA synthesis of an HCV E1 protein encoded by any one of the nucleic acid sequences shown in SEQ ID NOs:1-51 comprises:

(a) culturing a transformed or transfected host organism containing a nucleic acid sequence capable of directing the host organism to produce a protein, under conditions such that the protein is produced, said protein exhibiting substantial homology to a native E1 protein isolated from HCV having the amino acid sequence according to any one of the amino acid sequences shown in SEQ ID NOs:52-102 or combinations thereof.

In one embodiment, the RNA sequence of an HCV isolate was isolated and converted to cDNA as follows. Viral RNA is extracted from a biological sample collected from human subjects infected with hepatitis C and the viral RNA is then reverse transcribed and amplified by polymerase chain reaction using primers deduced from the sequence of HCV strain H-77 (Ogata et al. (1991)). Preferred primer sequences are shown as SEQ ID NOs:207-212 in the sequence listing. Once amplified, the PCR fragments are isolated by gel electrophoresis and sequenced.

In an alternative embodiment, the above method may be utilized for the recombinant DNA synthesis of an HCV core protein encoded by any one of the nucleic acid sequences shown in SEQ ID NOS: 103-154, where the protein produced by this method exhibits substantial homology to a native core protein isolated from HCV having amino acid sequence according to any one of the amino acid sequences shown in SEQ ID NOS: 155-206 or combinations thereof.

The vectors contemplated for use in the present invention include any vectors into which a nucleic acid sequence as described above can be inserted, along with any preferred or required operational elements, and which vector can then be subsequently transferred into a host organism and replicated in such organisms. Preferred vectors are those whose restriction sites have been well documented and which contain the operational elements preferred or required for transcription of the nucleic acid sequence.

The "operational elements" as discussed herein include at least one promoter, at least one operator, at least one leader sequence, at least one terminator codon, and any other DNA sequences necessary or preferred for appropriate transcription and subsequent translation of the vector nucleic acid. In particular, it is contemplated that such vectors will contain at least one origin of replication recognized by the host organism along with at least one selectable marker and at least one promoter sequence capable of initiating transcription of the nucleic acid sequence.

In construction of the recombinant expression vectors of the present invention, it should additionally be noted that multiple copies of the nucleic acid sequence of interest (either E1 or core) and its attendant operational elements may be inserted into each vector. In such an embodiment, the host organism would produce greater amounts per vector of the desired E1 or core protein. The number of multiple copies of the nucleic acid sequence which may be inserted into the vector is limited only by the ability of the resultant vector due to its size, to be transferred into and replicated and transcribed in an appropriate host microorganism.

Of course, those skilled in the art would readily understand that copies of both core and E1 nucleic acid sequence may be inserted into single vector such that a

host organism transformed or transfected with said vector would produce both the desired E1 and core proteins. For example, a polysistronic vector in which multiple different E1 and/or core proteins may be expressed from a single vector is created by placing expression of each protein under control of an internal ribosomal entry site (IRES)(Molla, A. et al. Nature, 356:255-257 (1992); Gong, S.K. et al. J. of Virol., 263:1651-1660 (1989)).

In another embodiment, restriction digest fragments containing a coding sequence for E1 or core proteins can be inserted into a suitable expression vector that functions in prokaryotic or eukaryotic cells. By suitable is meant that the vector is capable of carrying and expressing a complete nucleic acid sequence coding for an E1 or core protein. Preferred expression vectors are those that function in a eukaryotic cell. Examples of such vectors include but are not limited to vaccinia virus vectors, adenovirus or herpes viruses. A preferred vector is the baculovirus transfer vector, pBlueBac.

In yet another embodiment, the selected recombinant expression vector may then be transfected into a suitable eukaryotic cell system for purposes of expressing the recombinant protein. Such eukaryotic cell systems include but are not limited to cell lines such as HeLa, MRC-5 or CV-1. A preferred eukaryotic cell system is SF9 insect cells.

The expressed recombinant protein may be detected by methods known in the art including, but not limited to, Coomassie blue staining and Western blotting.

The present invention also relates to substantially purified and isolated recombinant E1 and core proteins. In one embodiment, the recombinant protein expressed by the SF9 cells can be obtained as a crude lysate or it can be purified by standard protein purification procedures known in the art which may include differential precipitation, molecular sieve chromatography,

ion-exchange chromatography, isoelectric focusing, gel electrophoresis and affinity and immunoaffinity chromatography. The recombinant protein may be purified by passage through a column containing a resin which has bound thereto antibodies specific for the open reading frame (ORF) protein.

The present invention further relates to the use of recombinant E1 and core proteins as diagnostic agents and vaccines. In one embodiment, the expressed recombinant proteins of this invention can be used in immunoassays for diagnosing or prognosing hepatitis C in a mammal. For the purposes of the present invention, "mammal" as used throughout the specification and claims, includes, but is not limited to humans, chimpanzees, other primates and the like. In a preferred embodiment, the immunoassay is useful in diagnosing hepatitis C infection in humans.

Immunoassays of the present invention may be those commonly used by those skilled in the art including, but not limited to, radioimmunoassay, Western blot assay, immunofluorescent assay, enzyme immunoassay, chemiluminescent assay, immunohistochemical assay, immunoprecipitation and the like. Standard techniques known in the art for ELISA are described in Methods in Immunodiagnosis, 2nd Edition, Rose and Bigazzi, eds., John Wiley and Sons, 1980 and Campbell et al., Methods of Immunology, W.A. Benjamin, Inc., 1964, both of which are incorporated herein by reference. Such assays may be a direct, indirect, competitive, or noncompetitive immunoassay as described in the art (Oellerich, M. 1984. J. Clin. Chem. Clin. BioChem 22:895-904) Biological samples appropriate for such detection assays include, but are not limited to serum, liver, saliva, lymphocytes or other mononuclear cells.

In a preferred embodiment, test serum is reacted with a solid phase reagent having surface-bound recombinant HCV E1 and/or core protein(s) as antigen(s). The solid

surface reagent can be prepared by known techniques for attaching protein to solid support material. These attachment methods include non-specific adsorption of the protein to the support or covalent attachment of the protein to a reactive group on the support. After reaction of the antigen with anti-HCV antibody, unbound serum components are removed by washing and the antigen-antibody complex is reacted with a secondary antibody such as labelled anti-human antibody. The label may be an enzyme which is detected by incubating the solid support in the presence of a suitable fluorimetric or calorimetric reagent. Other detectable labels may also be used, such as radiolabels or colloidal gold, and the like.

The HCV E1 and/or core proteins and analogs thereof may be prepared in the form of a kit, alone, or in combinations with other reagents such as secondary antibodies, for use in immunoassays.

In yet another embodiment the recombinant E1 and core proteins or analogs thereof can be used as a vaccine to protect mammals against challenge with hepatitis C. The vaccine, which acts as an immunogen, may be a cell, cell lysate from cells transfected with a recombinant expression vector or a culture supernatant containing the expressed protein. Alternatively, the immunogen is a partially or substantially purified recombinant protein. In yet another embodiment, the immunogen may be a fusion protein comprising core protein and a second, non-core protein joined together such that the core portion of the fusion protein will aggregate and "trap" the second protein on the surface of the particle produced by aggregation of the core protein. (Molecular Biology of the Hepatitis B Virus", McLachlan, A. (1991) CRC Press, Boca Raton, Fla.). Alternatively, the core protein could be mixed with the second protein <u>in vitro</u> to produce particles in which all or part of the second protein was exposed on the surface of the particle. Such particles would then serve as a carrier

in a multi-valent vaccine preparation. Second proteins or parts thereof which could be mixed with or fused to the core protein include, but are not limited to, HCV E1 and hepatitis B surface antigen.

While it is possible for the immunogen to be administered in a pure or substantially pure form, it is preferable to present it as a pharmaceutical composition, formulation or preparation.

The formulations of the present invention, both for veterinary and for human use, comprise an immunogen as described above, together with one or more pharmaceutically acceptable carriers and optionally other therapeutic ingredients. The carrier(s) must be "acceptable" in the sense of being compatible with the other ingredients of the formulation and not deleterious to the recipient thereof. The formulations may conveniently be presented in unit dosage form and may be prepared by any method well-known in the pharmaceutical art.

All methods include the step of bringing into association the active ingredient with the carrier which constitutes one or more accessory ingredients. In general, the formulations are prepared by uniformly and intimately bringing into association the active ingredient with liquid carriers or finely divided solid carriers or both, and then, if necessary, shaping the product into the desired formulation.

Formulations suitable for intravenous intramuscular, subcutaneous, or intraperitoneal administration conveniently comprise sterile aqueous solutions of the active ingredient with solutions which are preferably isotonic with the blood of the recipient. Such formulations may be conveniently prepared by dissolving the solid active ingredient in water containing physiologically compatible substances such as sodium chloride (e.g. 0.1-2.0m), glycine, and the like, and having a buffered pH compatible with physiological conditions to produce an

aqueous solution, and rendering said solution sterile. These may be present in unit or multi-dose containers, for example, sealed ampules or vials.

The formulations of the present invention may incorporate a stabilizer. Illustrative stabilizers are preferably incorporated in an amount of 0.10-10,000 parts by weight per part by weight of immunogens. If two or more stabilizers are to be used, their total amount is preferably within the range specified above. These stabilizers are used in aqueous solutions at the appropriate concentration and pH. The specific osmotic pressure of such aqueous solutions is generally in the range of 0.1-3.0 osmoles, preferably in the range of 0.8-1.2. The pH of the aqueous solution is adjusted to be within the range of 5.0-9.0, preferably within the range of 6-8. In formulating the immunogen of the present invention, an anti-adsorption agent may be used.

Additional pharmaceutical methods may be employed to control the duration of action. Controlled release preparations may be achieved through the use of polymer to complex or adsorb the proteins or their derivatives. The controlled delivery may be exercised by selecting appropriate macromolecules (for example polyester, polyamino acids, polyvinyl pyrrolidone, ethylenevinylacetate, methylcellulose, carboxymethylcellulose, or protamine sulfate) and the concentration of macromolecules as well as the methods of incorporation in order to control release. Another possible method to control the duration of action by controlled-release preparations is to incorporate the proteins, protein analogs or their functional derivatives, into particles of a polymeric material such as polyesters, polyamino acids, hydrogels, poly(lactic acid) or ethylene vinylacetate copolymers. Alternatively, instead of incorporating these agents into polymeric particles, it is possible to entrap these materials in microcapsules

prepared, for example, by coacervation techniques or by interfacial polymerization, for example, hydroxymethylcellulose or gelatin-microcapsules and poly (methylmethacylate) microcapsules, respectively, or in colloidal drug delivery systems, for example, liposomes, albumin microspheres, microemulsions, nanoparticles, and nanocapsules or in macroemulsions.

When oral preparations are desired, the compositions may be combined with typical carriers, such as lactose, sucrose, starch, talc, magnesium stearate, crystalline cellulose, methyl cellulose, carboxymethyl cellulose, glycerin, sodium alginate or gum arabic among others.

The E1 and core proteins of the present invention may also be used as a delivery system for anti-virals to prevent or attenuate HCV infection in a mammal by utilizing the property of both proteins to self-aggregate in vitro to "trap" the antiviral within the particles produced via aggregation of the core and E1 proteins. Examples of anti-virals which could be delivered by such a system include, but are not limited to antisense DNA or RNAs.

Vaccination can be conducted by conventional methods. For example, the immunogen or immunogens (e.g. the E1 protein may be administered alone or in combination with the E1 proteins derived from other isolates of HCV) can be used in a suitable diluent such as saline or water, or complete or incomplete adjuvants. Further, the immunogen(s) may or may not be bound to a carrier to make the protein(s) immunogenic. Examples of such carrier molecules include but are not limited to bovine serum albumin (BSA), keyhole limpet hemocyanin (KLH), tetanus toxoid, and the like. The immunogen(s) can be administered by any route appropriate for antibody production such as intravenous, intraperitoneal, intramuscular, subcutaneous, and the like. The immunogen(s) may be administered once or at periodic intervals until a significant titer of anti-HCV

antibody is produced. The antibody may be detected in the serum using an immunoassay.

In yet another embodiment, the immunogen may be nucleic acid sequence capable of directing host organism synthesis of E1 and/or core protein(s). Such nucleic acid sequence may be inserted into a suitable expression vector by methods known to those skilled in the art. Expression vectors suitable for producing high efficiency gene transfer in vivo include retroviral, adenoviral and vaccinia viral vectors. Operational elements of such expression vectors are disclosed previously in the present specification and are known to one skilled in the art. Such expression vectors can be administered intravenously, intramuscularly, subcutaneously, intraperitoneally or orally.

In an alternative embodiment, direct gene transfer may be accomplished via intramuscular injection of, for example, plasmid-based eukaryotic expression vectors containing a nucleic acid sequence capable of directing host organism synthesis of E1 and/or core protein(s). Such an approach has previously been utilized to produce the hepatitis B surface antigen in vivo and resulted in an antibody response to the surface antigen (Davis, H.L. et al. (1993) Human molecular Genetics, 2:1847-1851; see also Davis et al. (1993) Human Gene Therapy, 4:151-159 and 733-740).

Doses of E1 and/or core protein(s)-encoding nucleic acid sequence effective to elicit a protective antibody response against HCV infection range from about 1 to about 500 μg. A more preferred range being about 1 to about 500 μg.

The E1 and/or core proteins and expression vectors containing a nucleic acid sequence capable of directing host organism synthesis of E1 and/or core protein(s) may be supplied in the form of a kit, alone, or in the form of a pharmaceutical composition as described

above.

The administration of the immunogen(s) of the present invention may be for either a prophylactic or therapeutic purpose. When provided prophylactically, the immunogen(s) is provided in advance of any exposure to HCV or in advance of any symptom of any symptoms due to HCV infection. The prophylactic administration of the immunogen serves to prevent or attenuate any subsequent infection of HCV in a mammal. When provided therapeutically, the immunogen(s) is provided at (or shortly after) the onset of the infection or at the onset of any symptom of infection or disease caused by HCV. The therapeutic administration of the immunogen(s) serves to attenuate the infection or disease.

In addition to use as a vaccine, the compositions can be used to prepare antibodies to HCV E1 and core proteins. The antibodies can be used directly as antiviral agents or they may be used in immunoassays disclosed herein to detect HCV E1 and core proteins present in patient sera.. To prepare antibodies, a host animal is immunized using the E1 and/or core proteins native to the virus particle bound to a carrier as described above for vaccines. The host serum or plasma is collected following an appropriate time interval to provide a composition comprising antibodies reactive with the E1 or core protein of the virus particle. The gamma globulin fraction or the IgG antibodies can be obtained, for example, by use of saturated ammonium sulfate or DEAE Sephadex, or other techniques known to those skilled in the art. The antibodies are substantially free of many of the adverse side effects which may be associated with other anti-viral agents such as drugs.

The antibody compositions can be made even more compatible with the host system by minimizing potential adverse immune system responses. This is accomplished by removing all or a portion of the Fc portion of a foreign

species antibody or using an antibody of the same species as the host animal, for example, the use of antibodies from human/human hybridomas. Humanized antibodies (i.e., nonimmunogenic in a human) may be produced, for example, by replacing an immunogenic portion of an antibody with a corresponding, but nonimmunogenic portion (i.e., chimeric antibodies). Such chimeric antibodies may contain the reactive or antigen-binding portion of an antibody from one species and the Fc portion of an antibody (nonimmunogenic) from a different species. Examples of chimeric antibodies, include but are not limited to, non-human mammal-human chimeras, rodent-human chimeras, murine-human and rat-human chimeras (Robinson et al., International Patent Application 184,187; Taniguchi M., European Patent Application 171,496; Morrison et al., European Patent Application 173,494; Neuberger et al., PCT Application WO 86/01533; Cabilly et al., 1987 Proc. Natl. Acad. Sci. USA 84:3439; Nishimura et al., 1987 Canc. Res. 47:999; Wood et al., 1985 Nature 314:446; Shaw et al., 1988 J. Natl. Cancer Inst. 80:15553, all incorporated herein by reference).

General reviews of "humanized" chimeric antibodies are provided by Morrison S., 1985 Science 229:1202 and by Oi et al., 1986 BioTechniques 4:214.

Suitable "humanized" antibodies can be alternatively produced by CDR or CEA substitution (Jones et al., 1986 Nature 321:552; Verhoeyan et al., 1988 Science 239:1534; Biedleret al. 1988 J. Immunol. 141:4053, all incorporated herein by reference).

The antibodies or antigen binding fragments may also be produced by genetic engineering. The technology for expression of both heavy and light cain genes in E. coli is the subject of the PCT patent applications; publication number WO 901443, WO901443, and WO 9014424 and in Huse et al., 1989 Science 246:1275-1281.

The antibodies can also be used as a means of enhancing the immune response. The antibodies can be

administered in amount similar to those used for other therapeutic administrations of antibody. For example, normal immune globulin is administered at 0.02-0.1 ml/lb body weight during the early incubation period of other viral diseases such as rabies, measles, and hepatitis B to interfere with viral entry into cells. Thus, antibodies reactive with the HCV E1 and/or core proteins can be passively administered alone or in conjunction with another anti-viral agent to a host infected with an HCV to enhance the immune response and/or the effectiveness of an antiviral drug.

Alternatively, anti-HCV E1 antibodies and anti-HCV core antibodies can be induced by administered anti-idiotype antibodies as immunogens. Conveniently, a purified anti-HCV E1 or anti-HCV core antibody preparation prepared as described above is used to induce anti-idiotype antibody in a host animal, the composition is administered to the host animal in a suitable diluent. Following administration, usually repeated administration, the host produces anti-idiotype antibody. To eliminate an immunogenic response to the Fc region, antibodies produced by the same species as the host animal can be used or the Fc region of the administered antibodies can be removed. Following induction of anti-idiotype antibody in the host animal, serum or plasma is removed to provide an antibody composition. The composition can be purified as described above for anti-HCV E1 and anti-HCV core antibodies, or by affinity chromatography using anti-HCV E1 or anti-HCV core antibodies bound to the affinity matrix. The anti-idiotype antibodies produced are similar in conformation to the authentic HCV E1 or core protein and may be used to prepare an HCV vaccine rather than using an HCV E1 or core protein.

When used as a means of inducing anti-HCV virus antibodies in an animal, the manner of injecting the antibody is the same as for vaccination purposes, namely intramuscularly, intraperitoneally, subcutaneously or the

like in an effective concentration in a physiologically suitable diluent with or without adjuvant. One or more booster injections may be desirable.

The HCV E1 and core proteins of the invention are also intended for use in producing antiserum designed for pre- or post-exposure prophylaxis. Here an E1 or core protein, or mixture of E1 and/or core proteins is formulated with a suitable adjuvant and administered by injection to human volunteers, according to known methods for producing human antisera. Antibody response to the injected proteins is monitored, during a several-week period following immunization, by periodic serum sampling to detect the presence of anti-HCV E1 and/or anti-HCV core serum antibodies, using an immunoassay as described herein.

The antiserum from immunized individuals may be administered as a pre-exposure prophylactic measure for individuals who are at risk of contracting infection. The antiserum is also useful in treating an individual post-exposure, analogous to the use of high titer antiserum against hepatitis B virus for post-exposure prophylaxis.

For both <u>in vivo</u> use of antibodies to HCV virus-like particles and proteins and anti-idiotype antibodies and diagnostic use, it may be preferable to use monoclonal antibodies. Monoclonal anti-HCV E1 and anti-HCV core protein antibodies or anti-idiotype antibodies can be produced as follows. The spleen or lymphocytes from an immunized animal are removed and immortalized or used to prepare hybridomas by methods known to those skilled in the art. (Goding, J.W. 1983. Monoclonal Antibodies: Principles and Practice, Pladermic Press, Inc., NY, NY, pp. 56-97). To produce a human-human hybridoma, a human lymphocyte donor is selected. A donor known to be infected with HCV (where infection has been shown for example by the presence of anti-virus antibodies in the blood or by virus culture) may serve as a suitable lymphocyte donor. Lymphocytes can be isolated from a peripheral blood sample

or spleen cells may be used if the donor is subject to splenectomy. Epstein-Barr virus (EBV) can be used to immortalize human lymphocytes or a human fusion partner can be used to produce human-human hybridomas. Primary in vitro immunization with peptides can also be used in the generation of human monoclonal antibodies.

Antibodies secreted by the immortalized cells are screened to determine the clones that secrete antibodies of the desired specificity. For monoclonal anti-E1 and anti-core antibodies, the antibodies must bind to HCV E1 and core proteins respectively. For monoclonal anti-idiotype antibodies, the antibodies must bind to anti-E1 and anti-core protein antibodies respectively. Cells producing antibodies of the desired specificity are selected.

The present invention also relates to the use of single-stranded antisense poly- or oligonucleotides derived from nucleotide sequences substantially homologous to those shown in SEQ ID NOs:1-51 to inhibit the expression of hepatitis C E1 genes. The present invention further relates to the use of single-stranded anti-sense poly- or oligo-nucleotides derived from nucleotide sequences substantially homologous to those shown in SEQ ID NOs:103-154 to inhibit the expression of hepatitis C core genes. Alternatively, the anti-sense poly- or oligo-nucleotides may be complementary to both the E1 and core genes and hence, inhibit the expression of both hepatitis C E1 and core genes. By substantially homologous as used throughout the specification and claims to describe the nucleic acid sequences of the present invention, is meant a level of homology between the nucleic acid sequence and the SEQ ID NOs. referred to in the above sentence. Preferably, the level of homology is in excess of 80%, more preferably in excess of 90%, with a preferred nucleic acid sequence being in excess of 95% homologous with the DNA sequence shown in the indicated SEQ ID NO. These anti-sense poly- or oligonucleotides can be either DNA or RNA. The targeted

sequence is typically messenger RNA and more preferably, a single sequence required for processing or translation of the RNA. The anti-sense poly- or oligonucleotides can be conjugated to a polycation such as polylysine as disclosed in Lemaitre, M. et al. ((1989) Proc. Natl. Acad. Sci. USA 84:648-652) and this conjugate can be administrated to a mammal in an amount sufficient to hybridize to and inhibit the function of the messenger RNA.

The present invention further relates to multiple computer-generated alignments of the nucleotide and deduced amino acid sequences shown in SEQ ID NOs:1-206. Computer analysis of the nucleotide sequences shown in SEQ ID NOs:1-51 and 103-154 and of the deduced amino acid sequences shown in SEQ ID NOs:52-102 and 155-206 can be carried out using commercially available computer programs known to one skilled in the art.

In one embodiment, computer analysis of SEQ ID NOs:1-51 by the program GENALIGN (Intelligenetics, Inc. Mountainview, CA) results in distribution of the 51 HCV E1 sequences into twelve genotypes based upon the degree of variation of the sequences. For the purposes of the present invention, the nucleotide sequence identity of E1 cDNAs of HCV isolates of the same genotype is in the range of about 85% to about 100% whereas the identity of E1 cDNA sequences of different genotypes is in the range of about 50% to about 80%.

The grouping of SEQ ID NOs:1-51 into twelve HCV genotypes is shown below.

SEQ ID NOs:	Genotypes
1-8	I/1a
9-25	II/1b
26-29	III/2a
30-33	IV/2b
34	2c
35-39	V/3a
40	4a
41	4b
42-43	4c
44	4d
45-50	5a
51	6a

For those genotypes containing more than one E1 nucleotide sequence, computer alignment of the constituent nucleotide sequences of the genotype was conducted using GENALIGN in order to produce a consensus sequence for each genotype. These alignments and their resultant consensus sequences are shown in Figures 1A-G for the seven genotypes (I/1a, II/1b, III/2a, IV/2b, V/3a, 4c and 5a) which comprise more than one nucleotide sequence. Further alignment of the consensus sequences of Figures 1A-G with SEQ ID NO:34 (genotype 2c), SEQ ID NO:40 (genotype 4a), SEQ ID NO:41 (genotype 4b), SEQ ID NO:44 (genotype 4d) and SEQ ID NO:51 (genotype 6a) produces a consensus sequence for all twelve genotypes as shown in Figure 1H. The multiple alignments of nucleotide sequences shown in Figures 1A-H produce consensus sequences which serve to highlight regions of homology and non-homology between sequences found within the same genotype or in different genotypes and hence, these alignments can be used by one skilled in the art to design oligonucleotides useful as reagents in diagnostic assays for HCV.

Examples of purified and isolated oligonucleotide sequences derived from the consensus sequences shown in Figures 1A-H include, but are not limited to, SEQ ID NOs:213-239 where these oligonucleotides are useful as "genotype-specific" primers and probes since these

oligonucleotides can hybridize specifically to the nucleotide sequence of the E1 gene of HCV isolates belonging to a single genotype. The genotype-specificity of the oligonucleotides shown in SEQ ID NOs:213-239 is as follows: SEQ ID NOs:213-214 are specific for genotype I/1a; SEQ ID NOs:215-216 are specific for genotype II/1b; SEQ ID NOs:217-218 are specific for genotype III/2a; SEQ ID NOs:219-220 are specific for genotype IV/2b; SEQ ID NOs:221-223 are specific for genotype 2c; SEQ ID NOs:224-226 are specific for genotype V/3a; SEQ ID NOs:227-228 are specific for genotype 4a; SEQ ID NOs:229-230 are specific for genotype 4b; SEQ ID NOs:231-232 are specific for genotype 4c; SEQ ID NOs:233-234 are specific for genotype 4d; SEQ ID NOs:235-236 are specific for genotype 5a and SEQ ID NOs:237-239 are specific for genotype 6a.

In another embodiment, the computer analysis of SEQ ID NOs:103-154 by the program GENALIGN results in distribution of the 52 HCV core sequences into 14 genotypes based upon the degree of variation of the sequences.

The grouping of SEQ ID NOs:103-154 into 14 HCV genotypes is shown below.

SEQ ID NOs:	Genotypes
103-108	I/1a
109-124	II/1b
125-128	III/2a
129-133	IV/2b
134	2c
135-138	V/3a
139	4a
141	4b
143	4c
144	4c
145	4d
142	4e
140	4f
146-153	5a
154	6a

These 14 genotypes can be further grouped into 6

major genotypes designated genotypes 1-6 where genotype 1 comprises the sequences contained in minor genotypes I/1a and II/1b; genotype 2 comprises the sequences contained in minor genotypes III/2a, IV/2b and 2c; genotype 3 comprises sequences contained in genotype V/3a; genotype 4 comprises sequences contained in minor genotypes 4a-4f; genotype 5 comprises the sequences contained in genotype 5a and genotype 6 comprises the sequence contained in genotype 6a. Computer alignment of the constituent nucleotide sequences of the core cDNAs falling within genotypes I/1a, II/1b, III/2a, IV/2b, V/3a and 5a, to produce a consensus sequence for each of these genotypes is shown in Figures 6A (I/1a), 6B (II/1b), 6D (III/2a), 6E (IV/2b), 6G (V/3a) and 6I (5a). The alignment of the sequences found in minor genotypes I/1a and II/1b to produce a consensus sequence for major genotype 1 is shown in Figure 6C. The alignment of the sequences contained in minor genotypes III/2a, IV/2b and 2c to produce a consensus sequence for major genotype 2 is shown in Figure 6F. The alignment of the nucleotide sequences contained in minor genotypes 4a-4f to produce a consensus sequence for major genotype 4 is shown in Figure 6H. Further alignment of the consensus sequences shown in Figures 6C, 6F, 6G, 6H and 6I with SEQ ID NO:154 (genotype 6a/major genotype 6) to produce a consensus sequence for all genotypes is shown in Figure 6J and alignment of the consensus sequences shown in Figures 6A, 6B, 6D, 6E, 6G and 6I with 4a), SEQ ID NO:141 (genotype 4b), SEQ ID NO:143 (genotype 4c), SEQ ID NO:145 (genotype 4d), SEQ ID NO:142 (genotype 4e), SEQ ID NO:140 (genotype 4f) and SEQ ID NO:154 (genotype 6a) to produce a consensus sequence for all fourteen genotypes is shown in Figure 6K. As with the alignments of the envelope (E1) nucleotide sequences, the consensus sequences shown in Figures 6A-6K serve to highlight regions of homology and non-homology between sequences found within the same genotype or in different genotypes and hence, can be used by one skilled in the art

° to design oligonucleotides useful as reagents in diagnostic assays for HCV.

For example, purified and isolated oligonucleotide sequences derived from the consensus sequences shown in Figures 6A-6K may be useful as genotype-specific primers and probes since these oligonucleotides can hybridize specifically to the nucleotide sequence of the core gene of HCV isolates belonging to a given genotype. Examples of regions of the consensus sequence of the core gene of a given genotype from which primers specific for that genotype may be deduced include but are not limited to, the nucleotide domains shown below for each genotype. The sequence in which the indicated nucleotide domains are found are indicated in parentheses to the right of each genotype.

Genotype 1 (Consensus Sequence of Figure 6C)
427-466, 444-483, 447-486 (5'-3', sense)
505-466, 522-483, 525-486 (5'-3', antisense)

Genotype 1a (Consensus Sequence of Figure 6A)
141-180, 279-318 (5'-3', sense)
219-180, 246-207 (5'-3', antisense)

Genotype 1b (Consensus Sequence of Figure 6B)
67-106, 127-186, 234-273 (5'-3', sense)
144-106, 225-186, 311-272, 312-273 (5'-3', antisense)

Genotype 2 (Consensus Sequence of Figure 6F)
153-192, 162-201, 164-203, 168-207, 171-210, 182-221, 192-231, 193-232, 302-341 (5'-3', sense)
231-192, 240-201, 242-203, 246-207, 249-210, 260-221, 270-231, 271-232, 380-341 (5'-3', antisense)

Genotype III/2a (Consensus Sequence of Figure 6D)
276-315, 306-355 (5'-3', sense)
309-270, 354-315, 394-355, 571-532 (5'-3', antisense)

° Genotype IV/2b (Consensus Sequence of Figure 6E)
6-45, 135-174, 177-216, 309-348, 337-376, 375-414, 501-540 (5'-3', sense)
84-45, 213-174, 255-216, 387-348, 415-376, 453-414, 571-532, 573-540 (5'-3', antisense)

Genotype 2c (SEQ ID NO:134)
194-233, 273-312, 279-318, 417-456, 423-462, 504-543, 505-544, 517-556 (5'-3', sense)
272-233, 351-312, 354-315, 357-318, 450-411, 495-456, 501-462, 573-543, 556-573 (5'-3', antisense)

Genotype 3 or Genotype V/3a (Consensus Sequence of Figure 6G)
8-47, 45-84, 68-107, 87-126, 88-127, 90-129, 111-150, 142-181, 173-212, 177-216, 261-300,
276-315, 452-491, 520-559, 521-560, 529-568, 532-571, 533-572. (5'-3', sense)
86-47, 123-84, 146-107, 165-126, 186-147, 189-150, 219-180, 250-211, 251-212, 255-216,
339-300, 530-491, 573-543, 573-557, 573-559, 573-560. (5'-3', antisense)

Genotype 4 (Consensus Sequence of Figure 6H)
20-59 (5'-3', sense)
97-58, 98-59 (5'-3', antisense)

Genotype 4a (SEQ ID NO:139)
111-150, 150-189, 174-213, 183-222, 192-231, 261-300, 376-415, 396-435, 531-570 (5'-3', sense)
186-147, 252-213, 270 -231, 339-300, 454-415 (5'-3', antisense)

Genotype 4b (SEQ ID NO:141)
27-66, 30-69, 106-145, 271-310, 433-472, 447-486, 453-492 (5'-3', sense)

° 105-66, 183-144, 184-145, 345-306, 348-309, 349-310, 468-429, 510-471, 522-483, 570-531 (5'-3', antisense)

Genotype 4c (SEQ ID NO:143

174-213, 180-219, 207-246, 231-270 (5'-3', sense)

249-210, 252-213, 258-219, 309-270, 504-465 (5'-3', antisense)

Genotype 4d (SEQ ID NO:145)

173-212, 188-327, 430-469 (5'-3', sense)

248-209, 249-210, 250-211, 251-212, 366-327, 508-469 (5'-3', antisense)

Genotype 4e (SEQ ID NO:142)

160-199, 267-306, 287-326, 288-327, 524-564 (5'-3', sense)

238-199, 345-306, 365-326, 216-177, 522-483 (5'-3', antisense)

Genotype 4f (SEQ ID NO:140)

18-57, 36-75, 228-267, 396-435 (5'-3', sense)

96-57, 114-75, 306-267 (5'-3', antisense)

Genotype 5 or 5a (Consensus Sequence of Figure 6I)

176-215, 177-216, 181-220, 195-234, 221-260, 252-291, 255-294, 396-435, 435-474, 447-486, 498-537 (5'-3', sense)

254-215, 299-260, 310-271, 330-291, 333-294, 354-315, 464-425, 471-432, 483-444, 570-531 (5'-3', antisense)

Genotype 6 or 6a (SEQ ID NO:154)

20-59, 136-175, 156-195, 159-198, 175-214, 185-224, 277-316, 278-317, 312-351, 348-387,405-444, 406-445, 407-446, 408-447, 411-450, 432-471, 433-472, 435-474, 522-561 (5'-3', sense).

98-59, 214-175, 234-195, 237-198, 253-214, 262-223, 263-224, 354-315, 355-316, 382-343, 390-351, 426-387, 468-429, 483-444, 484-445, 485-446, 486-447, 489-450, 510-471, 511-

° 472, 513-474 (5'-3', antisense)

Such nucleotide domains may range from about 15 to about 100 bases in length with a more preferred range being about 30 to about 60 bases in length.

In an alternative embodiment, universal primers able to hybridize to the nucleotide sequences of the core gene of HCV isolates belonging to all of the genotypes disclosed herein may be deduced from universally conserved nucleotide domains of the consensus sequence shown in Figures 6J and 6K. Examples of such nucleotide domains include, but are not limited to, those shown below:

nucleotides 1-20, 1-25, 1-26, 1-27, 1-33, 50-89, 51-90, 52-91, 53-92, 61-100, 62-101, 77-116, 78-117, 79-118, 80-119, 81-120, 82-121, 83-122, 84-123, 85-124, 86-125, 97-136, 98-137, 99-138, 100-139, 101-140, 102-141, 329-368, 330-369, 331-370, 332-371, 354-393, 355-394, 356-395, 362-401, 363-402, 364-403, 365-404, 369-408, 442-481, 443-482, 457-496, 458-497, 475-514, 476-515, 477-516 (5'-3, sense); and

nucleotides 40-1, 41-2, 42-3, 43-4, 51-12, 52-13, 55-16, 56-17, 57-18, 58-19, 61-22, 62-23, 63-24, 64-25, 70-31, 124-85, 125-86, 126-87, 127-88, 128-89, 129-90, 136-97, 137-98, 138-99, 149-110, 150-111, 151-112, 152-113, 153-114, 154-115, 155-116, 156-117, 157-118, 158-119, 159-120, 170-131, 171-132, 172-133, 173-134, 174-135, 175-136, 403-364, 405-365, 406-366, 406-367, 430-391, 431-392, 432-393, 436-397, 437-398, 438-399, 439-400, 517-478, 518-479, 519-480, 532-493, 533-494, 550-511, 551-512 (5'-3', antisense)

Those skilled in the art would readily understand that the term "antisense" as used herein refers to primer sequences which are the complementary sequence of the indicated consensus sequence or SEQ ID NO:. Further, provided with the above examples of regions of the consensus sequences or indicated SEQ ID NOS: from which to deduce universal and genotype-specific primers, those

skilled in the art would readily be able to select pairs of primers, one sense and one antisense, which would be useful in the detection of HCV genotypes via the PCR methods described herein.

In yet another embodiment, the sequences shown in SEQ ID NO.:103-154 and the resultant consensus sequences produced by alignment of these SEQ ID NOs as shown in Figures 6A-6K may also be useful in the design of hybridization probes specific for a given HCV genotype. Examples of nucleotide domains of the consensus sequence or SEQ ID NO of a given genotype from which genotype-specific hybridization probes may be deduced include, but are not limited to, those shown below where the sequence from which the domains are found is indicated in parentheses to the right of each genotype.

Genotype	Position
1a (Consensus sequence of Figure 6A)	50-85 155-205 207-277 281-333 429-477 530-573
1b (Consensus sequence of Figure 6B)	81-131 159-225 252-318 411-472 530-573
2a (Consensus sequence of Figure 6D)	35-75 200-276 290-340 330-380 410-472 530-573
2b (Consensus sequence of Figure 6E)	20-70 149-199 191-241 240-285 261-318 323-373 351-401 389-439 429-477

	530-573
2c (SEQ ID NO:134)	208-258 230-276 290-345 411-460 430-490 530-573
3a (Consensus sequence of Figure 6G)	1-50 40-100 100-160 145-190 190-240 275-325 411-455 466-516 530-573
4a (SEQ ID NO:139)	35-85 145-195 200-250 255-305 341-390 390-440 530-573
4b (SEQ ID NO:141)	35-85 120-170 180-225 230-275 285-335 405-455 462-492 530-573
4c (SEQ ID NO:143)	35-85 190-246 245-295 282-318 372-415 440-480 530-573
4d (SEQ ID NO:145)	35-85 187-237 302-352 405-455 444-494 530-573
4e (SEQ ID NO:142)	35-85 57-84

	174-224 230-275 290-340 422-472 530-573
4f (SEQ ID NO:140)	35-85 174-224 242-292 290-340 422-472 530-573
5a (Consensus sequence of Figure 6I)	180-234 265-315 315-355 420-486 530-573
6a (SEQ ID NO:154)	34-84 150-200 180-230 230-290 291-333 341-395 429-490 530-573
1 (Consensus sequence of Figure 6C)	192-241 435-495
2 (Consensus sequence of Figure 6F)	186-240 320-360 440-475
4 (Consensus sequence of Figure 6H)	40-80

In yet another embodiment, universal hybridization probes may be derived from the consensus sequences shown in Figures 6J and 6K. Examples of nucleotide domains of the consensus sequences shown in Figure 6J and 6K from which universal hybridization probes may be derived include, but are not limited to, 1-33; 85-141; 364-408; 478-516.

The oligonucleotides of this invention can be synthesized using any of the known methods of oligonucleotide synthesis (e.g., the phosphodiester method of Agarwal et al. 1972, Agnew. Chem. Int. Ed. Engl. **11**:451,

the phosphotriester method of Hsiung et al. 1979, Nucleic Acids Res **6**:1371, or the automated diethylphosphoramidite method of Baeucage et al. 1981, Tetrahedron Letters **22**:1859-1862), or they can be isolated fragments of naturally occurring or cloned DNA. In addition, those skilled in the art would be aware that oligonucleotides can be synthesized by automated instruments sold by a variety of manufacturers or can be commercially custom ordered and prepared. In a preferred embodiment, the oligonucleotides of the present invention are synthetic oligonucleotides. The oligonucleotides of the present invention may range from about 15 to about 100 nucleotides; with the preferred sizes being about 20 to about 60 nucleotides; a more preferred size being about 25 to about 50 nucleotides; and a most preferred size being about 30 to about 40 nucleotides.

The present invention also relates to methods for detecting the presence of HCV in a mammal, said methods comprising analyzing the RNA of a mammal for the presence of hepatitis C virus.

The RNA to be analyzed can be isolated from serum, liver, saliva, lymphocytes or other mononuclear cells as viral RNA, whole cell RNA or as poly(A)$^+$ RNA. Whole cell RNA can be isolated by methods known to those skilled in the art. Such methods include extraction of RNA by differential precipitation (Birnbiom, H.C. (1988) Nucleic Acids Res., 16:1487-1497), extraction of RNA by organic solvents (Chomczynski, P. et al. (1987) Anal. Biochem., 162:156-159) and extraction of RNA with strong denaturants (Chirgwin, J.M. et al. (1979) Biochemistry, 18:5294-5299). Poly(A)$^+$ RNA can be selected from whole cell RNA by affinity chromatography on oligo-d(T) columns (Aviv, H. et al. (1972) Proc. Natl. Acad. Sci., 69:1408-1412). A preferred method of isolating RNA is extraction of viral RNA by the guanidinium-phenol-chloroform method of Bukh et al. (1992a).

The methods for analyzing the RNA for the presence of HCV include Northern blotting (Alwine, J.C. et al. (1977) Proc. Natl. Acad. Sci., 74:5350-5354), dot and slot hybridization (Kafatos, F.C. et al. (1979) Nucleic Acids Res., 7:1541-1522), filter hybridization (Hollander, M.C. et al. (1990) Biotechniques; 9:174-179), RNase protection (Sambrook, J. et al. (1989) in "Molecular Cloning, A Laboratory Manual", Cold Spring Harbor Press, Plainview, NY) and reverse-transcription polymerase chain reaction (RT-PCR) (Watson, J.D. et al. (1992) in "Recombinant DNA" Second Edition, W.H. Freeman and Company, New York).

A preferred method for analyzing the RNA is RT-PCR. In this method, the RNA can be reverse transcribed to first strand cDNA using a primer or primers derived from the nucleotide sequences shown in SEQ ID NOs:1-51 or SEQ ID NOs:103-154 or sequences complementary to those described. Once the cDNAs are synthesized, PCR amplification is carried out using pairs of primers designed to hybridize with sequences in the HCV E1 or core cDNA which are an appropriate distance apart (at least about 50 nucleotides) to permit amplification of the cDNA and subsequent detection of the amplification product. Alternatively, one can amplify both E1 and core cDNA sequences by using a primer pair where one primer hybridizes with the E1 cDNA sequence and the other primer hybridizes with the core cDNA sequence. Each primer of a pair is a single-stranded oligonucleotide of about 20 to about 60 bases in length with a more preferred range being about 30 to about 50 bases in length where one primer (the "upstream" primer) is complementary to the original RNA and the second primer (the "downstream" primer) is complementary to the first strand of cDNA generated by reverse transcription of the RNA. The target sequence is generally about 100 to about 300 base pairs long but can be as large as 500-1500 base pairs. Optimization of the amplification reaction to

obtain sufficiently specific hybridization to the nucleotide sequence of interest (either E1 or core or both E1 and core) is well within the skill in the art and is preferably achieved by adjusting the annealing temperature.

In one embodiment, the primer pairs selected to amplify E1 and core cDNAs are universal primers. By "universal", as used to describe primers throughout the claims and specification, is meant those primer pairs which can amplify E1 and/or core gene fragments derived from an HCV isolate belonging to any one of the genotypes of HCV described herein. Purified and isolated universal primers for E1 cDNAs are used in Example 1 of the present invention and are shown as SEQ ID NOs:207-212 where SEQ ID NOs:207 and 208 represent one pair of primers, SEQ ID NOs:209 and 210 represent a second pair of primers and SEQ ID NOs:211-212 represent a third pair of primers. Nucleotide domains of the consensus sequence shown in Figure 6J from which universal primers for core cDNAs may be deduced have previously been disclosed within the present specification. Alternatively, a universal primer for E1 cDNA sequence and a universal primer for core cDNA sequence may be used as a universal primer pair to amplify both E1 and core cDNAs.

In an alternative embodiment, primer pairs selected to amplify E1 and/or core cDNAs are genotype-specific primers. In the present invention, genotype-specific primer pairs can readily be derived from the following genotype-specific E1 nucleotide domains: nucleotides 197-238 and 450-480 of the consensus sequence of genotype I/1a shown in Figure 1A; nucleotides 197-238 and 450-480 of the consensus sequence of genotype II/1b shown in Figure 1B; nucleotides 199-238 and 438-480 of the consensus sequence of genotype III/2a shown in Figure C; nucleotides 124-177 and 450-480 of the consensus sequence of genotype IV/2b shown in Figure 1D; nucleotides 124-177, 193-238 and 436-480 of SEQ ID NO:34 (genotype 2C); nucleotides 168-207, 294-339 and 406-480 of the consensus

sequence of genotype V/3a shown in Figure 1E; nucleotides 145-183 and 439-480 of SEQ ID NO:40 (genotype 4a); nucleotides 168-207 and 432-480 of SEQ ID NO:41 (genotype 4b); nucleotides 130-183 and 450-480 of the consensus sequence of genotype 4c shown in Figure 1F; nucleotides 130-183 and 450-480 of SEQ ID NO:44 (genotype 4d); nucleotides 166-208 and 437-480 of the consensus sequence of genotype 5a shown in Figure 1b and nucleotides 168-207, 216-252 and 429-480 of SEQ ID NO:51 (genotype 6a). Genotype-specific HCV core nucleotide domains from which genotype-specific primers may be deduced have previously been described herein. Those skilled in the art would readily appreciate that in a pair of genotype-specific primers, each primer is derived from different nucleotide domains specific for a given genotype. Also, it is understood by those skilled in the art that each pair of primers comprises one primer which is complementary to the original viral RNA and the other which is complementary to the first strand of cDNA generated by reverse transcription of the viral RNA. For example, in a pair of genotype-specific primers for genotype 4b, one primer would have a nucleotide sequence derived from region 168-207 of SEQ ID NO:40 and the other primer would have a nucleotide sequence which is the complement of region 432-480 of SEQ ID NO:40. One skilled in the art would readily recognize that such genotype-specific domains would also be useful in designing oligonucleotides for use as genotype-specific hybridization probes. Indeed, genotype-specific hybridization probes deduced from the E1 and core sequences of the present invention have been previously disclosed herein.

The amplification products of PCR can be detected either directly or indirectly. In one embodiment, direct detection of the amplification products is carried out via labelling of primer pairs. Labels suitable for labelling the primers of the present invention are known to one skilled in the art and include radioactive labels, biotin,

avidin, enzymes and fluorescent molecules. The derived labels can be incorporated into the primers prior to performing the amplification reaction. A preferred labelling procedure utilizes radiolabeled ATP and T4 polynucleotide kinase (Sambrook, J. et al. (1989) in "Molecular Cloning, A Laboratory Manual", Cold Spring Harbor Press, Plainview, NY). Alternatively, the desired label can be incorporated into the primer extension products during the amplification reaction in the form of one or more labelled dNTPs. In the present invention, the labelled amplified PCR products can be detected by agarose gel electrophoresis followed by ethidum bromide staining and visualization under ultraviolet light or via direct sequencing of the PCR-products. Thus, in one embodiment, the present invention relates to a method for determining the genotype of a hepatitis C virus present in a mammal where said method comprises: amplifying RNA of a mammal via RT-PCR using labelled genotype-specific primers for the amplification step of the cDNA produced by reverse transcription.

In yet another embodiment, unlabelled amplification products can be detected via hybridization with labelled nucleic acid probes radioactively labelled or, labelled with biotin, in methods known to one skilled in the art such as dot and slot blot hybridization (Kafatos, F.C. et al. (1979) or filter hybridization (Hollander, M.C. et al. (1990)).

In one embodiment, the nucleic acid sequences used as probes are selected from, and substantially homologous to, SEQ ID NOs:1-51 and/or SEQ ID NOs:103-154. Such probes are useful as universal probes in that they can detect PCR-amplification products of E1 and/or core cDNAs of an HCV isolate belonging to any of the HCV genotypes disclosed herein. The size of these probes can range from about 200 to about 500 nucleotides. In an alternative embodiment, the sequence alignments shown in Figures 1A-1H

and 6A-6J may be used to design oligonucleotides useful as universal hybridization probes. Examples of core and envelope nucleotide domains from which such universal oligonucleotides may be deduced are disclosed herein.

In yet another embodiment, the present invention relates to a method for determining the genotype of a hepatitis C virus present in a mammal where said method comprises:

(a) amplifying RNA of a mammal via RT-PCR to produce amplification products;

(b) contacting said products with at least one genotype-specific oligonucleotide; and

(c) detecting complexes of said products which bind to said oligonucleotide(s).

In this method, one embodiment of said amplification step is carried out using the universal primers for E1 or core cDNAs as disclosed above. In step (b) of this method, the genotype-specific sequences used as probes may be deduced from the genotype-specific E1 and core nucleotide domains disclosed herein. These probes are useful in specifically detecting PCR-amplification products of E1 or core cDNAs of HCV isolates belonging to one of the HCV genotypes disclosed herein. In a preferred embodiment, these probes are used alone or in combination with other probes specific to the same genotype.

For example, a probe having a sequence according to SEQ ID NO:213 can be used alone or in combination with a probe having a sequence according to SEQ ID NO:214. The probes used in this method can range in size from about 15 to about 100 nucleotides with a more preferred range being about 30 to about 70 nucleotides. Such probes can be synthesized as described earlier.

In an alternative embodiment, the genotype of the amplification product of step (a) may be determined by using the nucleic acid sequences shown in SEQ ID NOs: 1-51 and 103-154 as probes (Delwart, E. et al. (1993)) Science,

262: 1257-1261). Probes utilized in the method of Delwart et al. may range in size from about 100 to about 1,000 nucleotides with a more preferred probe size being about 200 to about 800 base pairs and a most preferred probe size being about 300 to about 700 nucleotides.

The nucleic acid sequence used as a probe to detect PCR amplification products of the present invention can be labeled in single-stranded or double-stranded form. Labelling of the nucleic acid sequence can be carried out by techniques known to one skilled in the art. Such labelling techniques can include radiolabels and enzymes (Sambrook, J. et al. (1989) in "Molecular Cloning, A Laboratory Manual", Cold Spring Harbor Press, Plainview, New York). In addition, there are known non-radioactive techniques for signal amplification including methods for attaching chemical moieties to pyrimidine and purine rings (Dale, R.N.K. et al. (1973) Proc. Natl. Acad. Sci., 70:2238-2242; Heck, R.F. (1968) S. Am. Chem. Soc., 90:5518-5523), methods which allow detection by chemiluminescence (Barton, S.K. et al. (1992) J. Am. Chem. Soc., 114:8736-8740) and methods utilizing biotinylated nucleic acid probes (Johnson, T.K. et al. (1983) Anal. Biochem., 133:126-131; Erickson, P.F. et al. (1982) J. of Immunology Methods, 51:241-249; Matthaei, F.S. et al. (1986) Anal. Biochem., 157:123-128) and methods which allow detection by fluorescence using commercially available products.

The present invention also relates to computer analysis of the amino acid sequences shown in SEQ ID NOs:52-102 by the program GENALIGN. This analysis groups the 51 amino acid sequences shown in SEQ ID NOs:52-102 into twelve genotypes based upon the degree of variation of the amino acid sequences. For the purposes of the present invention, the amino acid sequence identity of E1 amino acid sequences of the same genotype ranges from about 85% to about 100% whereas the identity of E1 amino acid sequences of different genotypes ranges from about 45% to

about 80%.

The grouping of SEQ ID NOs:52-102 into twelve HCV genotypes is shown below:

SEQ ID NOs:	Genotypes
52-59	I/1a
60-76	II/1b
77-80	III/2a
81-84	IV/2b
85	2c
86-90	V/3a
91	4a
92	4b
93-94	4c
95	4d
96-101	5a
102	6a

For those genotypes containing more than one E1 amino acid sequence, computer alignment of the constituent sequences of each genotype was conducted using the computer program GENALIGN in order to produce a consensus sequence for each genotype. These alignments and their resultant consensus sequences are shown in Figures 2A-G for the seven genotypes (I/1a, II/1b, III/2a, IV/2b, V/3a, 4c and 5a) which comprise more than one sequence. Further alignment of the consensus sequences shown in Figures 2A-G with the amino acid sequences of SEQ ID NO:85 (genotype 2c); SEQ ID NO:91 (genotype 4a); SEQ ID NO:92 (genotype 4b); SEQ ID NO:95 (genotype 4d) and SEQ ID NO:102 (genotype 6a) to produce a consensus amino acid sequence for all twelve genotypes is shown in Figure 2H. The multiple alignment of E1 amino acid sequences shown in Figures 2A-H produces consensus sequences which serve to highlight regions of homology and non-homology between E1 amino acid sequences of the same genotype and of different genotypes and hence, these alignments can readily be used by those skilled in the art to design peptides useful in assays and vaccines for the diagnosis and prevention of HCV infection.

In another embodiment, the computer analysis of

SEQ ID NOS: 155-206 by the probe genome results in distribution of the 52 HCV core sequences into 14 genotypes based upon identification of genotype-specific amino acid sequences.

The grouping of SEQ ID NOS: 155-206 into 14 HCV genotypes is shown below:

SEQ ID NOS:	Genotypes
155-160	I/1a
161-176	II/1b
177-180	III/2a
181-185	IV/2b
186	2c
187-190	V/3a
191	4a
193	4b
195	4c
196	4c
197	4d
194	4e
192	4f
198-205	5a
206	6a

These fourteen genotypes can be further grouped into six major genotypes designated genotypes 1-6 as described earlier for the core nucleotide sequences of the present application. Computer alignment of the amino acid sequences disclosed in SEQ ID NOS: 155-206 are shown in figures 7A-7J. As with the multiple alignments of the E-1 amino acid sequences, the consensus sequences shown in figure 7A-7J serve to highlight regions of homology and nonhomology between core amino acid sequences of the same genotype and of different genotypes and hence, these alignments can readily be used by those skilled in the art to design peptides useful in assays and vaccines for the diagnosis and prevention of HCV infection.

Examples of purified and isolated peptides deduced from the alignments shown in Figures 2A-2H include, but are not limited to, SEQ ID NOs:240-263 wherein these peptides are derived from two regions of the amino acid

sequences shown in Figures 2A-H, amino acids 48-80 and amino acids 138-160. The peptides shown in SEQ ID NOs. 240-263 are useful as genotype-specific diagnostic reagents since they are capable of detecting an immune response specific to HCV isolates belonging to a single genotype. The genotype-specificity of the peptides shown in SEQ ID NOs:240-263 are as follows: SEQ ID NOs:240 and 252 are specific for genotype IV/2b; SEQ ID NOs:241 and 253 are specific for genotype 2c; SEQ ID NOs:242 and 254 are specific for genotype III/2a; SEQ ID NOs:243 and 255 are specific for genotype V/a; SEQ ID NOs:244 and 256 are specific for genotype II/1b; SEQ ID NOs:245 and 257 are specific for genotype I/1a; SEQ ID NOs:246 and 258 are specific for genotype 4a; SEQ ID NOs:247 and 259 are specific for genotype 4c; SEQ ID NOs:248 and 260 are specific for genotype 4d; SEQ ID NOs:249 and 261 are specific for genotype 4b; SEQ ID NOs:250 and 262 are specific for genotype 5a and SEQ ID NOs:251 and 263 are specific for genotype 6a. In SEQ ID NO:240, Xaa at position 22 is a residue of Ala or Thr, Xaa at position 24 is a residue of Val or Ile, Xaa at position 26 is a residue of Val or Met; in SEQ ID NO:242, Xaa at position 5 is a Ser or Thr residue, Xaa at position 11 is an Arg or Gln residue, Xaa at position 12 is an Arg or Gln residue; in SEQ ID NO:243, Xaa at position 3 is a Pro or Ser residue, Xaa at position 33 is a Leu or Met residue; in SEQ ID NO:244, Xaa at position 5 is a Thr or Ala residue, Xaa at position 13 is a Gly, Ala, Ser, Val or Thr residue, Xaa at position 14 is a Ser, Thr or Asn residue, Xaa at position 15 is a Val or Ile residue, Xaa at position 16 is a Pro or Ser residue, Xaa at position 18 is a Thr or Lys residue, Xaa at position 19 is a Thr or Ala residue, Xaa at position 22 is an Arg or His residue, Xaa at position 32 is an Ala, Val or Thr residue; in SEQ ID NO:245, Xaa at position 3 is an Ala or Pro residue, Xaa at position 4 is a Val or Met residue, Xaa at position 5 is a Thr or Ala residue, Xaa at

position 17 is a Thr or Ala residue, Xaa at position 18 is a Thr or Ala residue, Xaa at position 23 is a His or Tyr residue; in SEQ ID NO:247, Xaa at position 10 is a Val or Ala residue, Xaa at position 11 is a Ser or Pro residue, Xaa at position 18 is an Asp or Glu residue Xaa at position 20 is a Leu or Ile residue; in SEQ ID NO:250, Xaa at position 3 is a Gln or His residue, Xaa at position 12 is an Asn, Ser or Thr residue, Xaa at position 13 is a Leu or Phe residue, Xaa at position 23 is an Ala or Val residue; in SEQ ID NO:252, Xaa at position 16 is a Val or Ala residue, Xaa at position 18 is a Glu or Gln residue; in SEQ ID NO:254, Xaa at position 2 is an Ala or Thr residue, Xaa at position 4 is a Met or Leu residue, Xaa at position 9 is an Ala or Val residue, Xaa at position 17 is an Ile or Leu residue, Xaa at position 20 is an Ile or Val residue, Xaa at position 21 is a Ser or Gly residue; in SEQ ID NO:151, Xaa at position 9 is a Val or Ile residue, Xaa at position 16 is a Leu or Val residue, Xaa at position 20 is an Ile or Leu residue; in SEQ ID NO:256, Xaa at position 2 is an Ala or Thr residue, Xaa at position 6 is a Val or Leu residue, Xaa at position 12 is an Ile or Leu residue, Xaa at position 16 is a Val or Ile residue, Xaa at position 17 is a Val, Leu or Met residue, Xaa at position 19 is a Met or Val residue, Xaa at position 21 is an Ala or Thr residue; in SEQ ID NO:257, Xaa at position 2 is a Thr or Ala residue, Xaa at position 6 is a Val, Ile or Met residue, Xaa at position 12 is an Ile or Val residue, Xaa at position 16 is a Ile or Val residue; in SEQ ID NO:155, Xaa at position 5 is a Leu or Val residue, Xaa at position 21 is a Thr or Ala residue; in SEQ ID NO:262, Xaa at position 1 is a Thr or Ala residue, Xaa at position 5 is a Val or Leu residue, Xaa at position 9 is a Leu, Met or Val residue, Xaa at position 23 is a Gly or Ala residue.

Examples of core amino acid domains from which genotype-specific peptides may be deduced, include but are not limited to, those shown below where the sequence in

which the indicated domains are found is given in parentheses to the right of each genotype:

Genotype		Amino Acid Domains
1a	(consensus sequence of Figure 7A)	67-78
1b	(consensus sequence of Figure 7B)	67-78
2	(consensus sequence of Figure 7F)	66-81 110-119
2a	(consensus sequence of Figure 7D)	67-78 115-125
2b	(consensus sequence of Figure 7E)	67-78 123-133
2c	(SEQ ID NO:186)	67-78 75-81 184-191
3a	(consensus sequence of Figure 7G)	8-22 32-46 67-78 158-170 180-191
4	(consensus sequence of Figure 7H)	14-23
4a	(SEQ ID NO:191)	67-78
4b	(SEQ ID NO:193)	45-57 67-78
4c	(SEQ ID NO:195)	67-78
4d	(SEQ ID NO:197)	67-78
4e	(SEQ ID NO:194)	67-78
4f	(SEQ ID NO:192)	67-78
5a	(consensus sequence of Figure 7J)	67-78
6a	(SEQ ID NO:206)	67-78 101-108 144-155 157-163

Those skilled in the art would be aware that the peptides of the present invention or analogs thereof can be synthesized by automated instruments sold by a variety of manufacturers or can be commercially custom-ordered and prepared. The term analog has been described earlier in the specification and for purposes of describing the peptides of the present invention, analogs can further include branched, cyclic or other non-linear arrangements of the peptide sequences of the present invention.

Alternatively, peptides can be expressed from nucleic acid sequences where such sequences can be DNA, cDNA, RNA or any variant thereof which is capable of

directing protein synthesis. In one embodiment, restriction digest fragments containing a coding sequence for a peptide can be inserted into a suitable expression vector that functions in prokaryotic or eukaryotic cells. Such restriction digest fragments may be obtained from clones isolated from prokaryotic or eukaryotic sources which encode the peptide sequence.

Suitable expression vectors and methods of isolating clones encoding the peptide sequences of the present invention have previously been described. In yet another embodiment, an oligonucleotide capable of directing host organism synthesis of the given peptide may be synthesized and inserted into the expression vector.

The preferred size of the peptides of the present invention is from about 8 to about 100 amino acids in length when the peptides are chemically synthesized with a more preferred size being about 8 to about 30 amino acids and a most preferred size being about 10 to about 20 amino acids in length. For recombinantly expressed peptides, the size may range from about 20 to about 190 amino acids in length with a more preferred size being about 70 amino acids.

The present invention further relates to the use of genotype-specific peptides in methods of detecting antibodies against a specific genotype of HCV in biological samples. In one embodiment, at least one genotype-specific peptide deduced from a genotype-specific core or E1 amino acid domain may be used in any of immunoassays described herein to detect antibodies specific for a single genotype of HCV. In another embodiment, at least one genotype-specific peptide deduced from a genotype-specific core nucleotide domain and at least one genotype-specific peptide deduced from an E1 amino acid domain may be used in an immunoassay to detect antibodies against a single genotype of HCV. A preferred immunoassay is ELISA.

It is understood by those skilled in the art that

the diagnostic assays described herein using genotype-specific oligonucleotides or genotype-specific peptides can be useful in assisting one skilled in the art to choose a course of therapy for the HCV-infected individual.

In an alternative embodiment, a mixture of genotype-specific peptides can be used in an immunoassay to detect antibodies against multiple genotypes of HCV disclosed herein. For example, a mixture of genotype-specific peptides deduced from E1 amino acid sequences may comprise at least one peptide selected from SEQ ID NOs:244-245 and 256-257; one peptide selected from SEQ ID NOs:240, 242, 252 and 254; one peptide selected from SEQ ID NOs:246-249 and 258-261; one peptide selected from SEQ ID NOs:250 and 262; one peptide selected from SEQ ID NOs:243 and 255; one peptide selected from SEQ ID NOs:242 and 254 and one peptide selected from SEQ ID NOs:244 and 263. In a preferred embodiment, the peptides of the present invention can be used in an ELISA assay as described previously for recombinant E1 and core proteins.

In an alternative embodiment, the peptide(s) utilized in an immunoassay to detect all the genotypes of HCV disclosed herein may be a universal peptide deduced from universally conserved amino acid domains of the E1 or core proteins disclosed herein.

Examples of universally conserved core amino acid domains within the consensus sequence shown in Figure 7J from which universal peptides may be deduced include, but are not limited to amino acid domains 23-35, 53-66, 93-108, 122-138, 150-156, and 165-181 of the consensus sequence. Examples of universally conserved E1 amino acid domains within the HCV E1 protein are located within the consensus sequence for the 51 HCV E1 proteins shown in Figure 2H of the present application. Examples of universally conserved domains within the consensus sequence shown in Figure 2H include, but are not limited to, amino acid domains 10-20, 111-120, and 124-137 of the consensus sequence. The

universal peptides of the present invention may be used in an immunoassay to detect antibodies in patient sera specific for any of the genotypes of HCV disclosed herein.

The peptides of the present invention or analogs thereof may be prepared in the form of a kit, alone or in combinations with other reagents such as secondary antibodies, for use in immunoassay.

In another embodiment, the genotype-specific and universal peptides of the present invention may be used to produce antibodies that will react against HCV E1 or core proteins in immunoassays. In one embodiment, a genotype-specific E1 or core peptide can be used alone or in combination with other E1 or core peptides specific to the same genotype as immunogens to produce antibodies specific to HCV proteins of a single genotype.

In another embodiment, a mixture of peptides specific for different genotypes may be used to produce antibodies that will react with HCV proteins of any genotype disclosed herein. More preferably, antibodies reactive with HCV proteins of any genotype may be produced by immunizing an animal with universal peptide(s) of the present invention. Examples of immunoassays in which such antibodies could be utilized to detect HCV E1 and core proteins in biological samples include, but are not limited to, radioimmunoassays and ELISAs. Examples of biological samples in which HCV E1 and core proteins could be detected includes, but it is not limited to, serum, saliva and liver.

Of course, those skilled in the art would readily understand that the genotype-specific and universal peptides of the present invention and expression vectors containing nucleic acid sequence capable of directing host organism synthesis of these peptides could also be used as vaccines against hepatitis C. Formulations suitable for administering the peptide(s) and expression vectors of the present invention as immunogen, routes of administration,

pharmaceutical compositions comprising the peptides expression vectors and so forth are the same as those previously described for recombinant E1 and core proteins.

The genotype-specific and universal peptides of the present invention and expression vectors containing nucleic acid sequence capable of direct host organism synthesis of these peptides may also be supplied in the form of a kit, alone, or in the form of a pharmaceutical composition as described above for recombinant E1 and core proteins.

Any articles or patents referenced herein are incorporated by reference. The following examples illustrate various aspects of the invention but are in no way intended to limit the scope thereof.

MATERIALS

Serum used in these examples was obtained from 84 anti-HCV positive individuals who were previously found to be positive for HCV RNA in a cDNA PCR assay with primer set *a* from the 5' NC region of the HCV genome (Bukh, J. et al. (1992 (b)) Proc. Natl. Acad. Sci. USA 89:4942-4946). These samples were from 12 countries: Denmark (DK); Dominican Republic (DR); Germany (D); Hong Kong (HK); India (IND); Sardinia, Italy (S); Peru (P); South Africa (SA); Sweden (SW); Taiwan (T); United States (US); and Zaire (Z).

Example 1

Identification of the cDNA Sequence of the E1 Gene of 51 Isolates of HCV via RT-PCR Analysis of Viral RNA Using Universal Primers

Viral RNA was extracted from 100 μl of serum by the guanidinium-phenol-chloroform method and the final RNA solution was divided into 10 equal aliquots and stored at -80°C as described (Bukh, et al. (1992 (a)). The sequences of the synthetic oligonucleotides used in the RT-PCR assay, deduced from the sequence of HCV strain H-77 (Ogata, N. et al. (1991) Proc. Natl. Acad. Sci. USA 88:3392-3396), are shown as SEQ ID NOs:207-212. One aliquot of the final RNA solution, equivalent to 10 μl of serum, was used for cDNA synthesis that was performed in a 20 μl reaction mixture using avian myeloblastosis virus reverse transcriptase (Promega, Madison, WI) and SEQ ID NO:208 as a primer. The resulting cDNA was amplified in a "nested" PCR assay by *Taq* DNA polymerase (Amplitaq, Perkin-Elmer/Cetus) as described previously (Bukh et al. (1992a)) with primer set *e* (SEQ ID NOs:207-210). Precautions were taken to avoid contamination with exogenous HCV nucleic acid (Bukh et al. 1992a)), and negative controls (normal, uninfected serum) were interspersed between every test sample in both the RNA extraction and cDNA PCR procedures. No false positive results were observed in the analysis. In most instances,

amplified DNA (first or second PCR products) was reamplified with primers SEQ ID NO:211 and SEQ ID NO:212 prior to sequencing since these two primers contained EcoR1 sites which would facilitate future cloning of the E1 gene. Amplified DNA was purified by gel electrophoresis followed by glass-milk extraction (Geneclean, BIO 101, LaJolla, CA) and both strands were sequenced directly by the dideoxy-nucleotide chain termination method (Bachman, B. et al. (1990) Nucl. Acids Res. 18:1309)) with phage T7 DNA polymerase (Sequenase, United States Biochemicals, Cleveland, OH), [alpha ^{35}S]dATP (Amersham, Arlington Heights, IL) or [alpha ^{33}P] dATP (Amersham or DuPont, Wilmington, DE) and sequencing primers. RNA extracted from serum containing HCV strain H-77, previously sequenced by Ogata, N. et al. (1991), was amplified with primer set e (SEQ ID NOs:207-210) and sequenced in parallel as a control. The nucleotide sequences of the envelope 1 (E1) gene of all 51 HCV isolates are shown as SEQ ID NOs:1 - 51. In all 51 HCV isolates, the E1 gene was exactly 576 nucleotides in length and did not have any in-frame stop codons.

Example 2

Computer Analysis of the Nucleotide and Deduced Amino Acid Sequences of the E1 Gene of 51 HCV Isolates

Multiple computer-generated alignments of the nucleotide (SEQ ID NOs:1-51, Figures 1A-H) and deduced amino acid sequences (SEQ ID NOs:52-102, Figures 2A-H) of the cDNAs of the 51 HCV isolates constructed using the computer program GENALIGN (Miller, R.H. et al. (1990) Proc. Natl. Acad. Sci. USA 87:2057-2061) resulted in the 51 HCV isolates being divided into twelve genotypes based upon the degree of variation of the E1 gene sequence as shown in table 1.

Table 1. Percent nucleotide (nt) and amino acid (aa) sequence identify of the E1 gene among the 12 HCV genotypes.

	I/1a	II/1b	III/2a	IV/2b	2c	(V)/3a	4a	4b	4c	4d	5a	6a	nt:
	89.9-97.6	72.0-76.2	59.2-63.7	56.1-58.3	60.8-62.8	63.0-66.3	63.9-67.2	64.9-66.8	62.7-64.4	67.7-69.4	62.3-67.2	62.2-63.9	I/1a
aa:		88.9-97.9	58.3-62.2	53.8-57.5	60.1-61.5	63.9-67.2	60.9-63.7	63.4-65.8	61.6-65.1	63.0-65.5	62.2-66.5	61.6-63.0	II/1b
I/1a	91.1-98.4		88.0-91.3	69.1-71.0	72.7-73.6	58.0-60.8	61.5-62.7	58.9-60.4	59.7-63.4	58.7-61.3	56.6-60.8	55.0-56.8	III/2a
II/1b	75.5-80.7	90.1-97.9		92.7-95.0	67.5-68.9	56.3-58.3	58.9-60.8	56.4-57.6	57.1-59.9	57.5-59.0	53.5-56.6	53.6-55.2	IV/2b
III/2a	58.3-64.6	52.6-56.8	89.1-92.7		---	57.5-58.2	59.2	58.5	58.0-58.3	58.9	56.9-57.1	57.6	2c
IV/2b	54.2-56.8	51.0-54.2	69.3-72.9	93.8-96.4		93.8-99.1	64.4-65.3	62.7-64.1	60.9-62.5	62.3-63.9	61.8-64.4	58.0-58.9	(V)/3a
2c	56.3-60.4	52.6-55.7	74.5-77.1	67.7-69.8	---		---	74.8	75.5-78.0	74.8	62.8-64.6	62.0	4a
(V)/3a	64.1-68.8	66.7-70.8	54.7-58.9	54.2-56.8	52.1-53.6	94.3-98.4		---	74.0-74.8	72.0	63.9-64.6	62.7	4b
4a	69.3-73.4	64.6-67.2	62.0-63.0	58.9-60.4	58.3	66.1-68.8	---		90.1	77.6-78.6	62.7-64.8	63.0-64.4	4c
4b	66.7-69.3	66.1-70.3	53.6-56.3	52.1-53.1	53.6	62.0-64.6	76.0	---		---	64.4-66.1	64.1	4d
4c	66.1-72.9	64.6-69.3	55.2-61.5	54.2-58.3	54.7-58.3	63.0-65.6	77.1-81.3	79.2-80.2	89.6		90.1-95.7	60.6-63.2	5a
4d	73.4-75.5	66.7-70.3	56.3-58.9	55.2-55.7	54.2	63.5-64.6	78.1	77.6	82.8	---		---	6a
5a	66.1-73.4	64.1-70.3	52.6-57.3	50.5-53.1	54.2-56.3	60.4-64.1	67.2-68.2	65.1-67.2	67.7-71.4	69.3-71.4	92.7-97.4		
6a	64.6-65.6	62.5-65.6	49.0-51.0	49.0-50.5	50.5	57.8-58.9	66.1	62.5	66.1-67.2	66.7	62.0-63.5	---	

Nucleotide sequences analyzed in compiling the above table are shown in SEQ ID NOs:1-51 while the amino acid sequences analyzed are shown in SEQ ID NOs:52-102. The grouping of SEQ ID NOs: into genotypes is previously described in the specification.

The nucleotide and amino acid sequence identity of HCV isolates of the same genotype was in the range of 88.0-99.1% and 89.1-98.4%, respectively, whereas that of HCV isolates of different genotypes was in the range of 53.5-78.6% and 49.0-82.8%, respectively. The latter differences are similar to those found when comparing the envelope gene sequences of the various serotypes of the related flaviviruses, as well as other RNA viruses. When microheterogeneity in a sequence was observed, defined as more than one prominent nucleotide at a specific position, the nucleotide that was identical to that of the HCV prototype (HCV1, Choo et al. (1989)) was reported if possible. Alternatively, the nucleotide that was identical to the most closely related isolate is shown.

Analysis of the consensus sequence of the E1 protein of the 51 HCV isolates from this study demonstrated that a total of 60 (30.3%) of the 192 amino acids of the E1 protein were invariant among these isolates (Fig. 3). Most impressive, all 8 cysteine residues as well as 6 of 8 proline residues were invariant. The most abundant amino acids (e.g. alanine, valine and leucine) showed a very low degree of conservation. The consensus sequence of the E1 protein contained 5 potential N-linked glycosylation sites. Three sites at positions 209, 305 and 325 were maintained in all 51 HCV isolates. A site at position 196 was maintained in all isolates except the sole isolate of genotype 2c. Also, a site at position 234 was maintained in all isolates except one isolate of genotype I/1a, all four isolates of genotype IV/2b and the sole isolate of genotype 6a. Conversely, only genotype IV/2b isolates had a potential glycosylation site at position 233. Further analysis revealed a highly conserved amino acid domain (aa 302-328) in the E1 protein with 20 (74.1%) of 27 amino acids invariant among all 51 HCV isolates. It is possible that the 5′ and 3′ ends of this domain are conserved due to important cysteine residues and N-linked glycosylation

sites. The central sequence, 5'-GHRMAWDMM-3' (aa 315-323), may be conserved due to additional functional constraints on the protein structure. Finally, although the amino acid sequence surrounding the putative E1 protein cleavage site was variable, an amino acid doublet (GV) at position 380 was invariant among all HCV isolates.

A dendrogram of the genetic relatedness of the E1 protein of selected HCV isolates representing the 12 genotypes is shown in Fig. 4. This dendrogram was constructed using the program CLUSTAL (Higgins, D.G. et al. (1988) Gene, 73:237-244) and had a limit of 25 sequences. The scale showing percent identity was added based upon manual calculation. From the 51 HCV isolates for which the complete sequence of the E1 gene region was obtained, 25 isolates representing the twelve genotypes were selected for analysis. This dendrogram in combination with the analysis of the E1 gene sequence of 51 HCV isolates in Table 1 demonstrates extensive heterogeneity of this important gene.

The worldwide distribution of the 12 genotypes among 74 HCV isolates is depicted in Fig. 5. The complete E1 gene sequence was determined in 51 of these HCV isolates (SEQ ID NOs:1-51), including 8 isolates of genotype I/1a, 17 isolates of genotype II/1b and 26 isolates comprising genotypes III/2a, IV/2b, 2c, 3a, 4a-4d, 5a and 6a. In the remaining 23 isolates, all of genotypes I/1a and II/1b, the genotype assignment was based on a partial E1 gene sequence since they did not represent additional genotypes in any of the 12 countries. The number of isolates of a particular genotype is given in each of the 12 countries studied. Of the twelve genotypes, genotypes I/1a and II/1b were the most common accounting for 48 (65%) of the 74 isolates. Analysis of the E1 gene sequences available in the GenBank data base at the time of this study revealed that all 44 such sequences were of genotypes I/1a, II/1b, III/2a and IV/2b. Thus, based upon E1 gene analysis, 8 new genotypes

of HCV have been identified.

Also of interest, different HCV genotypes were frequently found in the same country, with the highest number of genotypes (five) being detected in Denmark. Of the twelve genotypes, genotypes I/1a, II/1b, III/2a, IV/2b and V/3a were widely distributed with genotype II/1b being identified in 11 of 12 countries studied (Zaire was the only exception). In addition, while genotypes I/1a and II/1b were predominant in the Americas, Europe and Asia, several new genotypes were predominant in Africa.

It was also found that genotypes I/1a, II/1b, III/2a, IV/2b and V/3a of HCV were widely distributed around the world, whereas genotypes 2c, 4a, 4b, 4d, 5a and 6a were identified only in discreet geographical regions. For example, the majority of isolates in South Africa comprised a new genotype (5a) and all isolates in Zaire comprised 3 new closely related genotypes (4a, 4b, 4c). These genotypes were not identified outside Africa.

Example 3

Identification of the cDNA Sequence Of The Core Gene Of 52 Isolates Of HCV

Viral RNA extraction, cDNA synthesis and "nested" PCR were carried out as in Example 1. For the cDNA PCR assay HCV-specific synthetic oligonucleotides deduced from previously determined sequences that flank the C gene were used. Amplified DNA was purified by gel electrophoresis followed by glass-milk extraction as described in Example 1 or by electroelution and both strands were sequenced directly. In 44 of the 52 HCV isolates studied the procedures for direct sequencing described in Example 1 were utilized. For a number of the HCV isolates confirmatory sequencing was performed with the Applied Biosystems 373A automated DNA sequencer and 8 HCV isolates of genotype I/1a or II/1b were sequenced exclusively by this method. All 73 negative control samples interspersed

among the test samples were negative for HCV RNA.

The amplified DNA fragment obtained in 50 of the 52 HCV isolates was specifically designed to overlap with previously obtained 5'NC sequences (Bukh et al. (1992b) Proc. Natl. Acad. Sci. U.S.A. 89:4942-4946) and with the E1 sequences disclosed herein at approximately 80 nucleotide positions each. A complete match was observed in 6033 of 6035 overlapping nucleotides. Two discrepancies were observed in isolate US6 at nt 552 (C and T) and nt 561 (C and T) respectively. This may have been due to microheterogeneity at these nucleotide positions, since the remaining overlapping sequence was unique for isolate US6. In addition, there were 3 confirmed instances of microheterogeneity: nt 33 in isolate SA11 (C,T and T), nt 36 in isolate S45 (A,C and A), and nt 552 in isolate P10 (C,T and T). Overall, the excellent agreement in these overlapping sequences in this study with the NC sequences disclosed in Bukh et al. and with the E1 sequences disclosed herein definitively ruled out contamination as a source of non-authentic HCV sequences. Furthermore, this analysis proved that the sequences obtained were from a single population, and not from different populations as could happen in mixed infections.

The core (C) gene was exactly 573 nucleotides in length in all 52 HCV isolates with an amino terminal start codon and no in-frame stop codons. Microheterogeneity was observed in 26 of the 52 HCV isolates at 0.2-1.4% of the 573 nucleotide positions of the C gene, and resulted in changes in 0.5-1.0% of the 191 predicted amino acids in 12 of these isolates. A multiple sequence alignment was performed and it showed that the nucleotide identities of the C gene among these HCV isolates were in the range of 79.4-99.0%. In order to compare the genetic relatedness of HCV isolates in different gene regions, phylogenetic trees of the C gene of all 52 HCV isolates and the E1 gene of 51 HCV isolates were constructed using the unweighted pair-

group method with arithmetic mean (Nei, M. (1987) Molecular Evolutionary Genetics (Columbia University Press, New York, N.Y., pp. 287-326) (Figure 8). In both dendrograms a division of the 45 HCV isolates from which C and E1 genes had been cloned into at least six major genetic groups (genotypes 1-6) and 12 minor genetic groups (genotypes I/1a, II/1b, III/2a, IV/2b, 2c, V/3a, 4a-4d, 5a, and 6a) was observed. It is noteworthy that a major division in genetic distance between HCV isolates of genotype 2 and those of the other genotypes in the phylogenetic analyses of both gene sequences was observed. Furthermore, the divergence of the minor genotypes within genotype 2 exhibited a degree of heterogeneity that is equivalent to that observed among the major genotypes. Analysis of the C gene from isolates Z5 and Z8, which had a unique 5' NC sequence (Bukh et al. (1992)) but from which the E1 gene could not be amplified, revealed that these isolates represented two additional genotypes. The designations 4e and 4f are assigned to these genotypes that have not been described previously. Overall, the present specification demonstrates that the genetic relatedness of HCV isolates is equivalent when analyzing the most conserved gene (C) and one of the most variable genes (E1) of the HCV genome, thereby providing strong evidence for the suggested division into major and minor genotypes.

Example 4

Computer Analysis of the Nucleotide and Deduced Amino Acid Sequences Of The Core Gene Of 52 HCV Isolates

In order to study further the heterogeneity of the C gene, a consensus sequence of the core gene from the 52 HCV isolates (Fig. 6J) was obtained. A total of 335 (58.5%) of the 573 nucleotides of the C gene were invariant among these HCV isolates. Nucleotides at the 1st and 2nd codon positions were invariant at 70.7% and 81.7% of these positions, respectively, while nucleotides at the 3rd

position were invariant at only 23.0% of such positions. Stretches of 6 or more invariant nucleotides were observed from nucleotides 1-8, 22-27, 85-92, 110-125, 131-141, 334-340, 364-371, 397-404, and 511-516 and may be suitable for anchoring primers for amplification of HCV RNA in cDNA PCR assays.

Genotype-specific nucleotide positions of the core gene of hepatitis C virus were also noted for each of the genotypes. These genotype-specific nucleotides are shown below where each genotype-specific nucleotide is given in parentheses next to the nucleotide position in which it is found.

Genotype 1: 460 (C), 466 (C), 483 (C), 486 (G).

Genotype I/1a: 180 (T).

Genotype II/1b: 106 (C), 273 (G).

Genotype 2: 192 (C), 201 (A), 203 (A), 207 (G), 210 (C), 221 (A), 231 (A), 232 (A), 341(A).

Genotype III/2a: 315 (C), 355 (G).

Genotype IV/2b: 45 (A), 174 (G), 216 (C), 348 (A), 376 (A), 414 (T).

Genotype 2c: 233 (G), 312 (C), 318 (A), 456 (C), 462 (G), 543 (C), 556 (T).

Genotype V/3a: 47 (T), 84 (A), 106 (G), 126 (A), 150 (T), 212 (G), 216 (A), 300 (A), 491 (T), 559 (C), 560 (A), 568 (G), 571 (A), 572 (G)

Genotype 4: 59 (T).

Genotype 4a: 213 (A), 231 (G), 415 (A).

Genotype 4b: 66 (G), 145 (G), 310 (A).

Genotype 4c: 213 (T), 219 (A), 270 (T).

Genotype 4d: 212 (T), 327 (G), 469 (C).

Genotype 4e: 199 (C), 306 (A), 326 (A).

Genotype 4f: 57 (T), 75 (A), 267 (A).

Genotype 5a: 291 (G), 294 (C).

Genotype 6a: 59 (C), 175 (A), 195 (A), 198 (A), 214 (C), 224 (A), 316 (C), 351 (G), 387 (G), 444-447 (GGCT), 450 (G), 471-472 (AA), 474 (C).

These genotype-specific nucleotides are of utility in designing the genotype-specific PCR primers and hybridization probes.

Finally, although the full length nucleic acid sequence of the C gene of isolates representing genotypes I/1a, II/1b, III/2a, IV/2b and V/3a have been reported by others, those of 9 of the 14 genotypes (i.e., 2c, 4a-4f, 5a and 6a) have not been reported previously. In sum, by aligning the consensus sequences of the major genotypes, the present application enables those skilled in the art to map universally conserved sequences as well as genotype-specific sequences of the C gene among 14 genotypes of HCV.

In order to study the heterogeneity of the deduced C protein, a multiple sequence alignment of the predicted amino acids for all 52 HCV isolates was performed, and a consensus sequence was obtained (Fig. 7J). The identities of the predicted 191 amino acids of the C protein among these HCV isolates were in the range of 85.3-100.0%. A total of 132 (69.1%) of the 191 amino acids of the C protein were invariant. The most prevalent amino acids in the consensus sequence were glycine (13.6%),

arginine (12.6%), proline (11.0%), and leucine (9.9%). The most conserved amino acids were tryptophan (5 of 5 amino acids invariant), aspartic acid (5 of 5 amino acids invariant), proline (19 of 21 amino acids invariant) and glycine (23 of 26 amino acids invariant). Previous analyses indicated that HCV is evolutionarily related to pestiviruses (Miller et al. (1990) Proc. Natl. Acad. Sci. U.S.A. 87:2057-2061). In this regard, it is of interest to note that the C proteins of both viruses have a high content of proline residues (Collette M.S. et al. (1988) Virology 165:200-208), which are likely to be important in maintaining the structure of this protein. As is characteristic for a protein that binds to nucleic acid, the C protein has conserved amino acids that are basic and positively charged, and these are capable of neutralizing the negative charge of the HCV RNA encapsidated by this protein (Rice, C.M. et al. (1986) in Togaviridae and Flaviviridae, eds Schleinger, S. & Schlensinger, M.J. (Plenum Press, New York, N.Y.) pp. 279-326). Specifically, over 16% of the amino acids in the consensus sequence of the C protein of HCV are arginine and lysine that are located primarily in three clusters (i.e., from amino acids 6-23, 39-74 and 101-121) (Shih, C.M. et al. (1993) J. Gen. Virol. 67:5823-5832) (Fig. 7J). The 10 arginine and lysine residues within amino acids 39-62 are invariant among all 52 HCV isolates, suggesting that this domain may represent an important RNA-binding site. The capsid proteins of the related flavi-and pestiviruses (Miller et al. (1990)) also have a high content of arginine and lysine (Rice et al. (1986); Collette et al. (1988). Although there are three major hydrophilic regions (i.e., amino acids 2-23, 39-74 and 101-121) that are conserved in all 52 HCV isolates, the remainder of the C protein is hydrophobic. Interestingly, one such highly conserved hydrophobic domain from aa 24-39 is flanked by proline residues. The hydrophobic domains are likely to be

involved in protein-protein and/or protein-RNA interactions during assembly of the nucleocapsid, as well as in interaction with the lipoprotein envelope, as has been suggested for flaviviruses (Rice et al. (1986)). Other significant observations are: *(i)* a cluster of 5 invariant tryptophan residues from aa 76-107; *(ii)* the lack of an N-linked glycosylation site (N-X-T/S); *(iii)* two potential nuclear localization signals (i.e., PRRGPR at amino acids 38-43 and PRGRRQP at amino acids 58-64) that are present in all 52 HCV isolates (Shih et al. (1993)); and *(iv)* a putative DNA-binding motif SPRG at amino acids 99-102, found in 51 of the 52 HCV isolates, with SP present in all 52 isolates. This study demonstrates that the C protein has features that are highly conserved among the various genotypes of HCV, and that are known to be characteristic of capsid proteins of other related viruses.

It should also be noted that the phylogenetic analysis of the amino acid sequence of the C proteins was not capable of resolving the minor groups within genotypes 1 and 4 because of the conservation of this protein (data not shown). Indeed, only a few type-specific amino acids were identified. One striking example was that isolates of genotype 4 have an additional methionine at position 20 that is specific for this major genetic group. Finally, the conservation of the sequences surrounding the cleavage site between the C and the E1 proteins of the different genotypes, which has been determined to be between amino acid 191 (alanine) and aa 192 (tyrosine) in HCV isolates of genotype 1 was analyzed (Hijikata, M., et al. (1991) *Proc. Natl. Acad. Sci. USA* 88:5547-5551). The C-terminal sequence of C is serine-alanine in all but one of the 48 HCV isolates comprising genotypes 1, 2, 4, 5 and 6. However, all 4 HCV isolates of genotype 3 in this study, as well as isolates of genotype 3 published previously (Okamoto, H., et al. (1993) *J. Gen. Virol.* 74:2385-2390, Stuyver, L., et al. (1993) *Biochem. Biophys. Res. Comm.*

192:635-641), contain alanine-serine at this position. Thus, studies will be needed to determine the C/E1 cleavage site in genotype 3 isolates. Overall, the present invention application discloses the mapping of universally conserved sequences, as well as genotype-specific sequences, of the C protein among 14 genotypes of HCV.

Implications of the mapping of universally conserved and genotype-specific core nucleotide and amino acid core sequences for diagnosis of HCV infection and for determination of HCV genotypes

Detection of antibodies directed against the HCV core protein is important in the diagnosis of HCV infection. The recombinant C22-3 protein, spanning amino acids 2-120 of the C gene, is a major component of the commercially available second-generation anti-HCV tests. Several studies have indicated that the three major hydrophilic regions of the C protein contain linear immunogenic epitopes (summarized in J. Clin. Microbiol, 30:1989-1994) (Sällberg, M. et al. (1992). For example, antibodies against synthetic peptides from amino acids 1-18, 51-68 and 101-118 were detected in infected patients (Sällberg, M. et al. (1992)). The present application demonstrates that, while these immunogenic regions are highly conserved, genotype-specific differences are observed at several amino acid positions that may influence the specificity and sensitivity of the serological tests. One such example is that a single amino acid substitution at amino acid 110 has been demonstrated to affect sero-reactivity (Sällberg, et al. (1992)). Despite the high degree of conservation in the immunodominant regions of the C protein among the different genotypes, it is possible that genetic heterogeneity of the C protein could lead to false negative results in current serological tests.

With respect to genotype analysis, several methods have been used to determine the genotype of HCV isolates without resorting to sequence analysis. These

include PCR followed by: (i) amplification with type-specific primers (Okamoto, H. et al. (1992) J. Gen. Virol., 73:673-679); (ii) determination of restriction-length polymorphism (Simmons, P. et al. (1993) J. Gen. Virol., 74:661-668); and (iii) specific hybridization (Stuyver, L. (1993) J. Gen. Virol., 74:1093-1102). The proposed methods have primarily been based on 5' NC and C sequences. Previous studies suggested that 5' NC-based genotyping systems would only be predictive of the major genetic groups of HCV (Bukh, J., et al. (1992) *Proc. Natl. Acad. Sci. USA* 89:4942- 4946, Bukh, J., et al. (1993) *Proc. Natl. Acad. Sci. USA* 90:8234-8238). The most widely used C-based genotype system has been the PCR assay with type-specific primers that was designed for distinguishing HCV isolates of genotypes I/1a, II/1b, III/2a, IV/2b and V/3a (Okamoto, H., et al. (1993) *J. Gen. Virol.* 74:2385-2390, Okamoto, H. et al. (1992) *J. Gen. Virol.* 73:673-679). Since this system was developed prior to the identification of genotypes 2c, 4a-4f, 5a and 6a there are significant limitations to this typing system. For example, the primers specific for genotype IV/2b (nt 270-251) are as highly conserved within isolates of genotype 4c and 6a as within the isolates of genotype IV/2b. Thus, this assay probably can not distinguish among these genotypes. Another C-based approach involves distinguishing between genotypes 1 and 2 by type-specific antibody responses (Machida et al (1992) Hepatology, 16:886-891). Synthetic peptides composed of amino acids 65-81 were found to be genotype-specific for genotypes 1 and 2 in ELISA assays. The present analysis of amino acid sequences demonstrated significant variation within isolates of genotypes 1 and 2. Thus it is likely that these peptides will not identify all isolates of genotypes 1 and 2. Furthermore, the peptide for genotype 1 was highly conserved within isolates of genotypes 3 and 4 and might detect antibodies against these genotypes as well. Finally, it should be pointed out that

most isolates of genotypes 3 and 4 had an identical amino acid sequence at positions 65-81.

Example 5

Detection by ELISA Based on Antigen from Insect Cells Expressing Complete E1 Or Core Protein

Expression of E1 or Core protein in SF9 cells. A cDNA (eg SEQ ID NO:1) encoding a complete E1 protein (eg SEQ ID NO:52) or a cDNA (eg SEQ ID NO:103) encoding a complete core protein (e.g. SEQ ID NO:155) is subcloned into pBlueBac - Transfer vector (Invitrogen) using standard subcloning procedures. The resultant recombinant expression vector is cotransfected into SF9 insect cells (Invitrogen) by the Ca precipitation method according to the Invitrogen protocol.

ELISA Based on Infected SF9 cells. 5 x 10^6 SF9 cells infected with the above-described recombinant expression vector are resuspended in 1 ml of 10 mM Tris-HCl, pH 7.5, 0.15M NaCl and are then frozen and thawed 3 times. 10 ul of this suspension is dissolved in 10 ml of carbonate buffer (pH 9.6) and used to cover one flexible microtiter assay plate (Falcon). Serum samples are diluted 1:20, 1:400 and 1:8000, or 1:100, 1:1000 and 1:10000. Blocking and washing solutions for use in the ELISA assay are PBS containing 10% fetal calf serum and 0.5% gelatin (blocking solution) and PBS with 0.05% Tween -20 (Sigma, St.Louis, MO) (washing solution). As a secondary antibody, peroxidase-conjugated goat IgG fraction to human IgG or horse radish peroxidase-labelled goat anti-Old or anti-New World monkey immunoglobulin is used. The results are determined by measuring the optical density (O.D.) at 405 nm.

To determine if insect cells-derived E1 or core protein representing genotype I/a of HCV could detect anti-HCV antibody in chimpanzees infected with genotype I/1a of HCV, three infected chimpanzees are examined. The serum of

° all 3 chimpanzees are found to seroconvert to anti-HCV.

Example 6

Use of the Complete E1 Protein as a Vaccine

Mammals are immunized with purified or partially purified E1 protein in an amount sufficient to stimulate the production of protective antibodies. The immunized mammals challenged with various genotypes of HCV are protected.

It is understood by one skilled in the art that the recombinant E1 protein used in the above vaccine can also be used in combination with other recombinant E1 proteins having an amino acid sequence shown in SEQ ID NOs:52-102. In addition, recombinant core proteins having an amino acid sequence shown in SEQ ID NOs:155-206 could also be used in the above vaccine, either alone, in combination with other recombinant core proteins of the present invention, or in combination with recombinant E1 proteins having an amino acid sequence shown in SEQ ID NOs:52-102.

Example 7

Determination of the Genotype of an HCV Isolate Via Hybridization of Genotype-Specific Oligonucleotides to RT-PCR Amplification Products.

Viral RNA is isolated from serum obtained from a mammal and is subjected to RT-PCR as in Example 1 or Example 3. Following amplification, the amplified DNA is purified as described in Example 1 or Example 3 and aliquots of 100 ul of amplification product are applied to dots on a nitrocellulose filter set in a dot blot apparatus. The dots are then cut into separate dots and each dot is hybridized to a ^{32}P-labelled oligonucleotide specific for a single genotype of HCV. The oligonucleotides to be used as hybridization probes are

deduced from the consensus sequences shown in Figures 1A-1H or 6A-6J or from the SEQ ID NOs: representing E1 or core sequences comprising genotypes 4a-4f, 2c and 6a.

Example 8

ELISA Based on Synthetic Peptides Derived From E1 cDNA Sequences

E1 peptide(s) specific for genotype I/1a is placed in 0.1% PBS buffer and 50ul of a 1mg/ml solution of peptide is used to cover each well of the microtiter assay plate. Serum samples from two mammals infected with genotype I/1a HCV and from one mammal infected with genotype 5a HCV are diluted as in Example 3 and the ELISA is carried out as in Example 3. Both mammals infected with genotype I HCV react positively with peptides while the mammal infected with genotype 5a HCV exhibits no reactivity. One skilled in the art would readily understand that in the above experiment, core peptides specific for genotype I/1a could be used in place of, or in combination with the E1 genotype-specific peptide(s).

Example 9

Use of E1 Peptides as a Vaccine

Since the E1 genotype-specific peptides of the present invention are derived from two variable regions in the complete E1 protein, there exists support for the use of these peptides as a vaccine to protect against a variety of HCV genotypes. Mammals are immunized with peptide(s) selected from SEQ ID NOs: 136-159 in an amount sufficient to stimulate production of protective antibodies. The immunized mammals challenged with various genotypes of HCV are protected. One skilled in the art would readily understand that genotype-specific core peptides of the present invention could also be used either alone, in combination with each other, or in combination with the

° genotype-specific E1 peptides, as a vaccine to protect against a variety of HCV genotypes. In addition, the above vaccines may also be formulated using the universal core and/or E1 peptides of the present invention.

Sub's new seq. list. per p.# 13